

# MODEL-BASED STATISTICAL ESTIMATION ALGORITHMS FOR FUNCTIONAL STRUCTURAL VIROLOGY

A Dissertation

Presented to the Faculty of the Graduate School  
of Cornell University

in Partial Fulfillment of the Requirements for the Degree of  
Doctor of Philosophy

by

Kang Wang

January 2012

© 2012 Kang Wang  
ALL RIGHTS RESERVED

# MODEL-BASED STATISTICAL ESTIMATION ALGORITHMS FOR FUNCTIONAL STRUCTURAL VIROLOGY

Kang Wang, Ph.D.

Cornell University 2012

A technique that is used widely in structural biology utilizing transmission electron microscopy (TEM) is single-particle cryo-EM, in which one projection image of a plunge-frozen specimen of multiple identical copies of a macromolecule, oriented randomly on a thin layer of vitreous ice, is acquired [88, 33]. The key assumption of each particle being identical allows one to combine projections of particles with the same orientation to increase SNR and to combine averaged projections of the structure at different views to create a high resolution 3D reconstruction of a given structure [33, 34]. This technique is especially attractive in studying viruses, which are relatively rigid and large macromolecular complexes that are made up of identical subunits arranged in a regular pattern [104, 43]. However, several properties of this technique limit its applicability in structural virology. These limitations are summarized below along with new experimental and computational methods that we help develop to address these issues. 1) Single-particle cryo-EM is an *in vitro* technique. To gain *in vivo* structural information, we utilize whole-cell cryo-electron tomography (CET), in which a tilt series of projection images of a single cell infected with virus particles is acquired and a 3-D tomogram is reconstructed from these projections. Together with pattern recognition techniques, we are able to study how a virus particle interacts with cellular components of its host cell during its lifecycle in a reliable way. 2) The requirement of identical copies of a structure prevents

the use of cryo-EM from studying viruses that are pleomorphic. Because 3D information of each particle is obtained in CET data, CET of purified particles is a powerful technique to analyze components of virus particles that are similar to each other in a collective way. One excellent example is the glycoprotein spikes of enveloped viruses. We develop a computational method that can be used to analyze subregions of CET data cubes in a reliable and efficient way. 3) Single-particle cryo-EM is a static imaging method that only provide snap-shots of structural states of a virus. By developing a physical model based on image statistics of cryo-EM data, we show how one can accurately predict the conformational dynamics of a virus structure based on its cryo-EM reconstruction and obtain information concerning the mechanism of how a virus functions. Hopefully, this set of tools will enable biologists to study viruses in a more comprehensive way.

## BIOGRAPHICAL SKETCH

Kang Wang was born in 1983, Guangzhou, China. He attended Zhushigang Elementary School in Guangzhou from 1989-1995. After attending the first year of junior high at Guangfu Middle School in Guangzhou, he moved to San Francisco, USA to live with his grandparents. He finish junior high at Benjamin Franklin Middle School in San Francisco and attended Abraham Lincoln High School in 1997. After graduating from Lincoln High in 2001, he attended U.C. Berkeley for his undergraduate education, majoring in Bioengineering, with a special emphasis on signal processing and image reconstruction. During his undergraduate training, Kang worked as a research assistant at Prof. Steven Brenner's biophysical laboratory for three years, concentrating on genome sequence alignment problems. Concurrently, Kang worked as an engineering research intern at Lawrence Berkeley National Laboratory (LBNL) to help develop a Trauma Patient Tracking System (TPTS) under the guidance of Prof. Thomas Budinger and Dr. Jonathan Maltz. In the year of 2005, he was promoted to engineering research associate at LBNL and continued his work on the TPTS system for another year before pursuing his graduate study at Cornell University.

This thesis document is dedicated to my beloved grandfather, Dr. Chen-en Wang, whose passion in medicine and education inspires me to pursue a career in both academic research and clinical medicine, and my dear grandmother, Dr. Yuan-Zhu Guo, for taking care of me during those difficult years when I struggled to adapt to live in the U. S.. I would also like to dedicate this thesis document to my family who have given me the freedom to pursue any career path I want and support throughout my life.

## ACKNOWLEDGEMENTS

I would like to thank the Department of Biomedical Engineering for providing funding for my first year at Cornell, so I can truly experience the exciting and diverse research programs in the department. I am also indebted to my advisor, Prof. Peter C. Doerschuk, for his guidance throughout my thesis research and valuable advice in career planning. In addition, I consider it a great honor to work with Dr. Fu Chi-yu and Prof. Jack Johnson at The Scripps Research Institute. Finally, I would like to thank members of my research lab, Charlene, Emma, John, Nathan, Ipek, Seunghee and Yili for their good companionship through my five years at Cornell and their advices and help throughout my thesis project.

## TABLE OF CONTENTS

Biographical Sketch . . . . .	iii
Dedication . . . . .	iv
Acknowledgements . . . . .	v
Table of Contents . . . . .	vi
List of Tables . . . . .	ix
List of Figures . . . . .	x
<b>1 Introduction</b>	<b>1</b>
1.1 Cryo-Electron Tomography (CET) of Purified Particles . . . . .	2
1.2 Whole Cell Cryo-Electron Tomography: Imaging Viruses <i>in situ</i> . . . . .	5
1.3 Molecular Dynamics based on Coarse-Grain Physical Model . . . . .	7
1.4 Statistical Estimation via Mathematical Model . . . . .	9
1.5 Outlook . . . . .	10
<b>2 A maximum-likelihood approach to the classification and alignment of CET cubes</b>	<b>12</b>
2.1 Introduction . . . . .	12
2.2 Mathematical model for the object . . . . .	15
2.3 3-D cube formation model . . . . .	17
2.4 Statistical estimation . . . . .	20
2.4.1 Maximum likelihood estimation . . . . .	20
2.4.2 Practical issues . . . . .	22
2.5 Synthetic data from the x-ray crystallographic structure of Hong Kong 97 . . . . .	23
2.5.1 Performance as a function of SNR and tilt range microscopy parameters . . . . .	24
2.5.2 Performance as a function of number of data cubes . . . . .	28
2.6 Experimental data from STIV infected <i>Sulfolobus sulfataricus</i> cells . . . . .	31
2.7 Discussion and conclusion . . . . .	33
<b>3 In Vivo Assembly of an Archaeal Virus Studied with Whole-Cell Cryo-Electron Tomography</b>	<b>43</b>
3.1 Introduction . . . . .	43
3.2 Assembly and Maturation of STIV Observed In Vivo . . . . .	46
3.3 Distribution and Packing of Viral Particles In Vivo . . . . .	48
3.4 Discussion . . . . .	49
3.5 Experimental Procedures . . . . .	52
3.5.1 Sample Preparation . . . . .	52
3.5.2 Tomography Data Collection and Image Processing . . . . .	53
3.5.3 Maximum Likelihood Reconstruction . . . . .	54
3.5.4 Template Matching Approach . . . . .	55



<b>4</b>	<b>Asymmetry Detection within Symmetric Structures</b>	<b>63</b>
4.1	Introduction . . . . .	63
4.2	Maximum Likelihood Estimator (MLE) based on Normalized Correlation . . . . .	64
4.3	Asymmetry Detection Algorithm . . . . .	66
4.4	Numerical Results . . . . .	70
4.4.1	Simulation Study with P22 . . . . .	70
4.4.2	Lambda Portal Detection and Asymmetric Reconstruction . . . . .	71
4.4.3	STIV Turrets Analysis . . . . .	73
4.4.4	Discussion . . . . .	74
<b>5</b>	<b>Efficient Computation of Maximum Likelihood Estimator (MLE) via Spherical Fourier Transform</b>	<b>82</b>
5.1	Introduction . . . . .	82
5.2	Maximum Likelihood Formulation of the Cryo-ET Problem . . . . .	83
5.3	Practical Implementation . . . . .	87
<b>6</b>	<b>Understanding conformational dynamics of macromolecular complexes from the heterogeneity of cryo-EM data</b>	<b>90</b>
6.1	Introduction . . . . .	90
6.2	Mechanical model . . . . .	93
6.3	Electron scattering intensity model . . . . .	95
6.4	Statistical mechanics . . . . .	97
6.5	Implications of Sections 6.2–6.4 . . . . .	97
6.5.1	Estimates of the mean and the auto-correlation of the electron scattering intensity from the data . . . . .	101
6.6	Estimating the mechanical properties from the data . . . . .	102
6.7	A model for $\mathcal{V}(\delta)$ and $V$ . . . . .	105
6.8	Characteristic frequencies and normal modes . . . . .	107
6.9	Practical Computation . . . . .	109
6.9.1	Equilibrium Positions Determination from Discretizing the continuous 3D Intensity . . . . .	109
6.9.2	Strategies for Computing the Eigensystem of $V$ . . . . .	111
6.10	Feasibility of Model Parameter Estimation . . . . .	114
6.10.1	The IgG model . . . . .	114
6.10.2	Model parameter estimation . . . . .	116
6.10.3	Inferring structural dynamics: NMA analysis . . . . .	117
6.11	Validation of Structural Dynamic Prediction Based on NMA Analysis . . . . .	118
6.12	A Real 3D Example: Modeling the Flock House Virus Capsid . . . . .	119
<b>A</b>	<b>Tomogram processing algorithms for Chapter 3</b>	<b>130</b>
A.1	Class Membership, Location, and Orientation Determination of a Particle in a CET Tomogram . . . . .	130

A.2	Determination of a lattice . . . . .	131
A.3	Orientational heterogeneity of the particles in a lattice . . . . .	133
<b>B</b>	<b>Spherical Fourier Transform and SO(3) Fourier Transform</b>	<b>134</b>
B.1	Representing the electron scattering intensity . . . . .	134
B.1.1	Symbolic . . . . .	134
B.1.2	Computation . . . . .	139
B.2	Correlation . . . . .	142
B.2.1	Symbolic . . . . .	142
B.2.2	Computation . . . . .	145
B.3	Representing $c_{l,m}(r)$ and $C_{l,m}(k)$ for computation . . . . .	146
B.4	Quantities related to correlation . . . . .	148
	<b>Bibliography</b>	<b>152</b>

## LIST OF TABLES

2.1	Resolution in nm at which that the pair of reconstructed structures are consistent with the pair of reference structures using a cutoff value of 0.5 in FSC. For each experimental condition (SNR and tilt range), a pair of values are reported, which are the resolutions of the reconstructions of the procapsid and capsid structures, respectively. . . . .	37
2.2	Number of alignment errors. For each experimental condition (SNR and tilt range), the entry is the number of the 60 data cubes in which the orientation estimate based on the reconstruction is not correct. . . . .	37
2.3	Size of alignment errors at SNR = 0.005 as a function of the range of tilt angles for those cubes that are misaligned. The error is quantified as the average of the error present in each pair of true and estimated orientations. The error in a single pair is quantified as the “Euclidean” or “Frobenius” matrix norm [68, p. 358] (the square root of the sum of the squares of the elements in the matrix) of the difference between the rotation matrices which correspond to the true and estimated orientations. Since the particle has icosahedral symmetry, there are 60 equivalent rotational matrices for any orientation and the choice which gives the lowest alignment error is the matrix that is used. . . . .	37
2.4	Summary of the second set of experiments with synthetic HK97 CET cubes. Resolutions in nm for the reconstructions of procapsid and capsid are reported in pairs in Row 1 for procapsid and capsid, respectively. Numbers of misaligned and misclassified cubes are reported in Rows 2 and 3, respectively. . . . .	41
3.1	The analysis of viral distribution and array packing . . . . .	57

## LIST OF FIGURES

2.1	Whole cell CET of <i>S. solfataricus</i> infected with STIV (a) A 10 nm slice of a 3-D tomogram, computationally sectioned perpendicular to the direction of beam. Multiple pyramid-like protrusions were observed. DNA-free procapsid and DNA-filled virions were clearly identified in the cytoplasm. The quasi-crystalline viral array was labeled. (b) & (c) Enlarged views of some breakage sites in the ruptured cells. The fragments of reminiscent pyramid structures were labeled, which lack S-layer and can be distinguished from typical cell wall. SL, S-layer; CM, cytoplasmic membrane; Pyr, pyramid like protrusion; Pyr Frag, pyramid fragments; IPB, intra-pyramidal body. Scale bar, 200 nm. . . . .	36
2.2	Reconstructions of full and empty particles from two of the four tomograms. Panels (a) and (b): First tomogram. Panels (c) and (d): Second tomogram. . . . .	38
2.3	Fourier Shell Correlation (FSC) curves between all pairs of full-particle reconstructions and empty-particle reconstructions. . . .	39
2.4	Single particle cryo EM reconstruction of STIV from Ref. [58]. . .	39
2.5	3-D surface rendering by Chimera of the low-resolution reference structures for HK97. Panels (a) and (b) show the mature capsid and the procapsid, respectively. . . . .	40
2.6	Cross-sections of simulated data subcubes of HK97 at various SNR settings and tilt ranges. Each row represent data from same tilt range, which are $\pm 70^\circ$ , $\pm 60^\circ$ , and $\pm 50^\circ$ from top to bottom. The interval between tilts is always $2^\circ$ . Each column represent data from same SNR setting, which are 0.1, 0.05, 0.01, and 0.005 from left to right. . . . .	40
2.7	Cross-sections in reciprocal space of the simulated data subcubes of HK97 for three different tilt ranges. Panels (a), (b), and (c) show tilt ranges of $\pm 70^\circ$ , $\pm 60^\circ$ , and $\pm 50^\circ$ , respectively. . . . .	41
2.8	Plots of the Fourier Shell Correlation results for correlation between reconstruction and reference structures (colored curves) and for correlation between data cubes and reference structures (black curves). The reconstructions are based on 10, 20, 30, and 60 synthetic HK97 cubes. Each cube is the result of simulating a single-tilt experiment with tilts in the range $\pm 60^\circ$ and with SNR = 0.01. The resolution gain achieved by this algorithm is essentially the difference between the black curves and the colored curves. . . . .	42

3.1	Whole-cell CET of <i>S. solfataricus</i> infected with STIV: (A) A 20 nm slice of a 3D tomogram, computationally sectioned perpendicular to the direction of the beam (B) Enlarged views of some representative pyramids. (C) Surface representations of a pyramid viewed from the side and the top of the structure. (D) Enlarged views of some pyramids at early stages of formation. (E) A model of pyramid formation. A pyramid forms by either mechanically protruding and/or enzymatically digesting through the S-layer. The S-layer structure is perturbed and finally detaches from pyramid membrane. Specific proteins and lipids are recruited as a pyramid builds. SL, S-layer; CM, cytoplasmic membrane; Pyr, pyramid like protrusion; STIV, STIV particles; IPB, intrapyramidal body. Scale bar, 200 nm (A), 100 nm (B and D). . . . .	56
3.2	STIV Virions, Procapsids, and Partially Assembled Particles Observed In Vivo Subtomographic slices displaying STIV virions (A) and procapsids (B) and partially assembled particles (C) that contain parts of capsid and membrane viewed perpendicular to the direction of the beam (x-y plane). Scale bar, 100 nm (C). . . .	57
3.3	The Reconstructed Density of the Two Classes of STIV Particles Determined by the ML Algorithm The reconstructions of a virion (A) and a procapsid (B) determined with 123 particle-containing subvolumes of the cellular tomogram. (C) The surface representation of the single molecule reconstruction of purified virions. (D) The radial density plots of the subtomographic reconstructions of the virion (blue), procapsid (red), and the single molecule reconstruction of purified virions (black). . . . .	58
3.4	The Analysis of Viral Distribution and Packing: A gallery of model representations of viral distributions with cell periphery outlined in gray. . . . .	59
3.5	The Analysis of Viral Distribution and Packing: The model representation of a quasicrystalline packing of a viral array (virions, blue icosahedra; procapsids, red icosahedra) . . . . .	60
3.6	(a) The box plot of the distance between the centers of the particles in clusters to its 1st to 6th closest neighboring particles. Red bar, median; Blue box, range covered by 25-75% of particles; Error bar, minimum and maximum in the distribution. (b) The histogram of the distance between the turrets of its neighboring particles within 75 nm . . . . .	60
3.7	The diagrams of unit cells that describe the packing of observed viral arrays . . . . .	61

3.8	A Model of STIV Assembly and Maturation The lipids and trans-membrane proteins assemble to form viral membrane as the trans-membrane proteins serve as tape-measure scaffolding and facilitate correct assembly of capsid proteins. The growing membrane and capsid form a procapsid. The genome is packaged through either a turret vertex or a specialized portal vertex. A procapsid matures to a virion without undergoing large-scale capsid transformation. . . . .	62
4.1	Results based on 210 experimental CET cubes and 3500 experimental cryo EM images of bacteriophage $\lambda$ . Panel (a) Example CET cube. Panels (b–c): Average of CET cubes aligned by which 5-fold axis is labeled as having the portal by the maximum likelihood subcube reconstruction algorithm. Panels (d–e): Average over symmetries of (b–c). Panels (f–g): Portal (6 fold) from (d–e). Panels (h–i): Pentamer (5-fold) from (d–e). Panel (j): Example cryo EM image. Panels (k–l): Asymmetric reconstruction from cryo EM images where (k) and (l) have the common (11 of 12) and the rare (1 of 12) 5-fold axis directed out of the page at the center of the image, respectively, and (m) shows the portal in blue on the right hand side. Different visualization orientations are used in (b–e,m) versus (k–l). (Visualization: UCSF Chimera [89]. Color: distance from origin). . . . .	78
4.2	Results with 63 experimental CET cubes of purified STIV. Panel (a): Example CET cube. Panel (b): The icosahedrally-symmetric maximum likelihood reconstruction [Step (2)(a)]. Panels (c–f): Reconstructions of Class 1 [(c–d)] and Class 2 [(e–f)]. Each subcube receives a class label and a rotational orientation [Step (4)]. After rotation to an absolute orientation, the subcubes sharing a class label are averaged and the result is further averaged over the rotational symmetry [5-fold for (c–d) and 6-fold for (e–f)] to compute (c–f). (Visualization: UCSF Chimera [89]. Color: distance from origin). . . . .	79
4.3	Asymmetric Detection Algorithm Float Chart Part (A) . . . . .	80
4.4	Asymmetric Detection Algorithm Float Chart Part (B) . . . . .	81
6.1	IgG model represented as spheres and rods. Each pink sphere represent one point mass of the model, connected by springs, represent as yellow rods with width proportional to the spring constant values. The spring constants are uniform everywhere in the IgG model except at the joints between the stem and the two arms which has a smaller spring constant value to allow the two arms to flex. . . . .	121
6.2	Two instances of the model. Open circles are the equilibrium positions and the filled circles are the instantaneous positions . .	122

6.3	Mean and variance of the image pixel values of the IgG model .	122
6.4	Mean and variance of the image pixel values of the IgG model .	123
6.5	Percentage error in the estimation of the model parameters (spring constant) of the IgG model. . . . .	123
6.6	A plot of the mass-spring model of adenylate kinase. The equilibrium positions, which are represented by circles, are obtained from discretizing the density map of adenylate kinase which is generated synthetically from its x-ray crystallographic structure. A line connecting two circles represents a spring connecting two point masses. . . . .	124
6.7	Variance of electron scattering intensity of adenylate kinase as predicted from the model. The surface of the adenylate structure is colored according to the variance information at that location. Red represents highest variance while blue represents the least variance Panel (a): A side view of the surface of adenylate kinase, same orientation as the spring-mass model plot. Panel (b): A top-down view of the surface of the adenylate kinase structure	125
6.8	The open/close forms of adenylate kinase. Panel (a): The first normal mode of the 3-D mass-spring model from the 3-D adenylate kinase density map. Panel (b): Displacement vectors between the open and close form of the x-ray crystallographic structure of adenylate kinase. The similarity of Panels (a) and (b) demonstrates the relevance of normal modes, and the underlying mechanical model, to the functioning of adenylate kinase.	126
6.9	Panel (a): The mean intensity reconstruction of FHV based on 528 cryo-EM experimental images. Panel (b): Mechanical model based on a discretization of the mean intensity shown in (a). Pink spheres represent the point masses and yellow rods represent connecting springs between two neighboring point masses. . . .	127
6.10	Panel (a): Experimental covariance matrix between the 50 point masses from one asymmetric unit of the FHV model. Panel (b): Predicted covariance matrix between the 50 point masses from one asymmetric unit of the FHV model. The spring parameters are the result of the last step in the minimization process, though the predicted mean and covariance never get reasonably close to the experimental mean and covariance. . . . .	128

6.11	A comparison between two Covariance matrix from the FHV model in which the spring constants are randomly chosen between the values of 0 to 100. It can be seen that the overall structure of the covariance matrix is similar even though the actual values might be different. This hints at that the spring connectivity actually places a strong constraint on the Covariance matrix structure. Panel (a): Predicted covariance matrix between the 50 point masses from one asymmetric unit of the FHV model with random spring constant. Panel (b): Predicted covariance matrix between the 50 point masses from one asymmetric unit of the FHV model with another set of random spring constant. . . . .	129
B.1	Conversions among $\rho(\mathbf{x})$ , $c_{l,m}(r)$ , $P(\mathbf{k})$ , and $C_{l,m}(k)$ . . . . .	136



## CHAPTER 1

### INTRODUCTION

Transmission Electron Microscopy (TEM) is an excellent tool to study structures of large macromolecular complexes, since such an electron dense biological specimen creates an excellent contrast under an electron microscope [43]. Because electron inelastic scattering will damage fragile biological specimen, electron dose for imaging biological specimen is typically kept at very low level, which results in images that are extremely noisy [111]. A technique that is used widely in structural biology utilizing TEM is single-particle cryo-EM, in which one projection image of a plunge-frozen specimen of multiple identical copies of a macromolecule, oriented randomly on a thin layer of vitreous ice, is acquired [88, 33]. The key assumption of each particle being identical allows one to combine projections of particles with the same orientation together to increase SNR and to combine averaged projections of the structure at different views to create a high resolution 3D reconstruction of a given structure [33, 34]. This technique is especially attractive in studying viruses, which are relatively rigid and large macromolecular complexes that are made up of identical subunits arranged in a regular pattern [104, 43]. Thus, a single projection image of a virus particle already provides a tremendous amount of structural information about the virus. In fact, for viruses with helical symmetry, all one needs is a single projection image and the three dimensional information of each subunit of the virus is obtained. Other viruses such as Flock House Virus (FHV) have capsids that possess icosahedral symmetry, which essentially increases the structural information obtained from one particle by 60-fold [111]. Due to these favorable conditions in viruses, their structures are resolved routinely in the sub-nanometer range, and recent advance in both experimental and computational

methods have led to structures of viruses resolved at near atomic resolution (i.e.,  $3\text{\AA} \sim 4\text{\AA}$ ) [43].

However, several properties of this technique limits its applicability in structural virology: 1) The requirement for a homogeneous set of particles limits the range of viruses that can be studied using this technique since many viruses have pleomorphic structural components. 2) Single particle cryo-EM is an *in vitro* technique that requires purification, hence no information can be obtained on how viruses interact with their host cell. 3) Single particle cryo-EM is a static technique in which the specimen are flash frozen before imaging so only snapshots of the structural states of a virus are obtained, which might not be sufficient for understanding how viruses function mechanistically.

In this thesis document, we describe several new experimental and computational methods that supplement structural information gained from single particle cryo-EM analysis, and hopefully will provide biologists with a set of powerful tools to study viruses in a more comprehensive way.

## **1.1 Cryo-Electron Tomography (CET) of Purified Particles**

The requirement for identical copies of a structure limits the range of viruses can be studied using single particle cryo-EM. For example, many enveloped viruses contain an amorphous envelop, in which glycoproteins, important in host-cell recognition and cell surface binding, are embedded [104, 30]. Due to lack of a homogeneous structure, no information on these envelops and their embedded glycoproteins can be obtained by single particle cryo-EM. Owing to the advance in automatic image data acquisition, an emerging technique is

electron tomography on frozen hydrated specimens (CET), in which a series of projection images are acquired by tilting the specimen at different angles and a 3D tomogram is reconstructed based on these projection images [74]. The added third dimension allows the study of subregions of each particle with ease since subregion information can be extracted easily from each CET data cube in contrast to the single-particle cryo-EM case in which all the information along the z-axis is added together in a projection image.

Such an advantage can be exploited in several ways in studying virus structures. First, CET can be used to study viruses that exhibit some form of pleomorphism, but with subregions that are relatively homogeneous [75]. Traditional ideas developed in cryo-EM analysis, such as classification and averaging, can be applied to these relatively homogeneous regions since they can be extracted readily from each 3D data cube from CET tomograms. This is exactly the case for enveloped viruses. Even though the envelop is amorphous, the envelope spikes, important in host cell recognitions are similar in structures from instance to instance, thus, they can be extracted out for further analysis collectively. In the case of identical glycoprotein structures, as in the Moloney Murine Leukemia Virus (MoMuLV), simple averaging after subregion extraction identified by template matching provides a 30Å structure using CET [30]. In another study using a similar processing method, it is found that glycoproteins form ordered patches through lateral interactions, which creates membrane curvatures that might drive membrane budding [50]. Perhaps the more interesting studies are those with spikes that exhibit some form of heterogeneity, as in the case of HIV spikes. The resulting class averages of HIV spikes are different among different groups, leading the authors in Ref. [128] to conclude that reference bias during classification might play a role in the discrepancy. Thus, it is important

to develop computational methods that can accurately sort through a set of heterogeneous data into relative homogeneous subsets reliably.

Another situation in which CET is useful is in the structural study of asymmetric regions within a highly symmetric virus structure, such as the portal structure at a 5-fold axis of a bacteriophage, which is important for understanding genome packaging. Such a small asymmetric region, unless electron dense, is usually not detectable in projection images, and thus not reveal from reconstruction of single particle cryo-EM. Again owing to the added third dimension, such regions are easier to detect in CET data. Recently, three groups have investigated the portal structures of herpesviruses by CET, all confirming that there is a unique portal at one specific 5-fold axis that is structurally different than the pentamers at the other eleven sites, but the precise location of the portal with respect to the capsid is inconsistent among the three groups [20, 23, 18]. This again shows the importance of a reliable computation method in structural studies. Our contribution is along the line of the second category of problems, in which we try to detect an asymmetric region within a highly symmetric virus structure. Unlike the previous studies, in which they either used a reference structure or experimental modifications to enhance the signals of the portal in order to successfully locate the specific location of the asymmetric region among the potential sites (i.e., one of twelve 5-fold axes), the algorithm presented in Chapter 4 only requires one to know the potential sites of asymmetry and the asymmetric structure is determined that is consistent with the set of data cubes from CET and no extra information is required, which makes the algorithm more general since such prior information might not be available and, even if available, would possibly bias the final reconstruction of such an asymmetric structure.

## 1.2 Whole Cell Cryo-Electron Tomography: Imaging Viruses *in situ*

Both cryo-EM and CET of purified particles are *in vitro* techniques that require lysis of the cell and subsequent purification to extract the virus particles. While such methods provide valuable information concerning virus capsid assemblies and maturation, they do not provide structural information concerning how viruses interact with the host cell in events such as virus entry, replication, assembly, and egress from the host cell [52]. Since viruses are obligate parasites, it is important to follow these events in the context of the host cell [35]. Whole-cell CET is another ET variant that allows the visualization of locations and dynamics of viruses with other structures and organelles in the host cell [104, 52]. Typically, a cell infected with the viruses of interest is cryo-preserved and ET is applied to obtain 3D information concerning the whole cell. One major limitation in imaging a cell is its thickness might exceed the maximum penetration depth of electron microscopy, which is around 1  $\mu\text{m}$ . Thus, whole cell CET is typically limited to small prokaryotic cells. In the case the specimen is too thick, it can be sectioned into thin slices using a microtome (diamond knives) before image acquisition. [82].

With whole-cell CET, many interesting supramolecular structures, which only exist within an intact cell, can be visualized during the life cycle of a virus [82]. For example, complexes of HIV/SIV particles docking with host cell membrane right before virus entry is visualized using CET in Ref. [104]. It is shown that distinct structures, termed "entry claws" which consist of clusters of closely packed rod-shaped like structures, exist between the contact region of virus par-

ticles and host cell membrane. [103]. Another interesting event in the virus life cycle that can be visualized with whole cell CET is how virus particles are transported within a cell. It is shown in Ref. [114] that Rice Gall Dwarf Virus (RGDV) particles aligned parallel on top of microtubules (MT), which suggest that they are conveyed via the MT-mediated system [52, 114].

A key in visualizing these events inside a host cell is a reliable way to identify the locations of virus particles. However, except for large continuous structures such as cell membranes and the actin filament network, it is generally hard to visualize macromolecular complexes such as virus particles within a cell tomogram due to its typically low SNR. [82] Pattern recognition techniques can be used to identify the locations of macromolecular complexes of interest within a cell tomogram objectively and reliably [82]. Cross-correlation based template matching is one such technique that can identify known structures within a cell tomogram robustly in a low SNR environment [10]. Template matching requires a reference, which might not be available for transient structures that only exists within a cell. In Chapter 2, we present a computational method that is able to sort through a set of heterogeneous data sets and reconstruct the sets of distinct structures that exist among the data at a higher resolution than the tomogram itself and without any initial references. This method is very useful in identify different transient structural states of a virus that might exist within an infected cell. By using these high-resolution references, a template matching technique can be used to identify their locations within the cell and provide insight into how viruses in different stages of the virus lifecycle interact within the host cell. In Chapter3, this technique is applied to analyze CET tomograms of an Archaea cell infected with the STIV virus. Using our classification and reconstruction technique, two structural states of STIV have been identified, and through sub-

sequent template matching with these two references, the locations of these two structural states of STIV within the cell are obtained. By analyzing these locations in a quantitative way, we gain insightful information concerning virus replication and packaging within a host cell.

A currently practiced goal of whole-cell CET is not to obtain a high-quality structure of the virus particles, but rather to study the spatial arrangement of virus particles inside the host cell and their spatial relationships with other cellular components of the host cell in order to understand the mechanism of virus docking, entry, replication, transportation and egress, each of which might require specific computational tools that are tailored to answer specific biological questions concerning the virus life cycle [44, 52].

### **1.3 Molecular Dynamics based on Coarse-Grain Physical Model**

All techniques using EM are essentially static methods, which provide a snapshot of the structural states of a virus particle, but do not provide structural dynamic information concerning virus particles, except in situations where intermediate states can be stimulated under specific experimental conditions, such as that in [115], where pH controls the maturation of HK 97 through several intermediate states, in which dynamics can be inferred based on these stable intermediates. However, such dynamics is often important in understanding how a macromolecular assembly functions.

Traditionally, studying molecular dynamics is accomplished by building a

physical model and simulating the various atomic forces interacting between atoms of the structure in order to predict how a structure "moves". This is a classical problem in protein folding in which all atoms of a structure are simulated based on interactions by various known inter-atomic forces and a feasible pathway for the molecule's movement is the one that leads to the lowest internal energy of the system. Such a model, though extremely accurate, quickly becomes computational infeasible when used to model large macromolecular complexes such as a virus [60]. Thus, in recent years, coarse-grained models such as the elastic network model based on blocks of atoms are used, which are shown to be effective in predicting major cohesive motions of macromolecular structures [92, 107]. In [59, 60], Kim et. al. provide an efficient method that is able to generate feasible pathways between stable structures of several important enzyme structures in which experimental data is able to validate its accuracy. Using this method, they are able to generate a pathway of HK97 virus capsid maturation that matches important rigid body rotations in two of its compact domains [61]. The accuracy of such a model depends critically on the construction of the model which depends on the springs that connect the point masses. In all current applications of the network model, spring constants of such parameters are assumed to be uniform and the same from structure to structure and are determined so that they match experimental data for a few specific structures where x-ray crystallographic data is known. We believe that such a model can be improved by assigning model parameters on a case by case basis and by allowing spring constants to be different to reflect the interactions between neighboring point masses. In Chapter 6, we present a method to estimate model parameters based on image statistics from single-particle cryo-EM data and validate the model by comparing the major motion predicted from our model against dis-



placement vectors of two forms of x-ray crystallographic structure of adenylyl kinase. Single-particle cryo-EM technique is essentially an averaging technique that captures the information concerning an ensemble of virus particles.

## **1.4 Statistical Estimation via Mathematical Model**

At the core of each computational method discussed above is a statistical estimation problem of some desired parameters that are related to the experimental data via a detail mathematical model. We choose to formulate the problem as a statistical estimation problem because typically the imaging data from EM is extremely noisy, and statistical methods are able to capture such uncertainty in a quantitative way, thus it provides us a powerful tool to assess the quality of data in a objective way. Perhaps even more powerful in statistical estimation is its ability to incorporate such uncertainty information into the estimation of the desired parameters, in which larger uncertainty of a particular piece of data will generally lead to its lesser contribution to the determination of the desired parameters, thus hedging against potential error that might arise. The advantage of a detail mathematical model relating desired parameters and experimental data is that prior knowledge from other experiments or known facts can be incorporated into the model in a way that constraints the solution to ensure that a reasonable result is obtained, which makes it robust even in an extremely low signal situation.

## 1.5 Outlook

As mentioned at the beginning of this introduction, the experimental and computational methods discussed in this thesis document are meant to complement the structural study of viruses by single-particle cryo-EM, and provide different, rather than superior, information. The best approach is to combine these available methods and utilize the strength of each. For example, in Chapter 4, we show how one can reliably detect and reconstruct asymmetric regions within a highly symmetric structure based on CET cubes. However, the reconstruction from CET cubes generally has lower resolution than reconstruction from cryo-EM data. Thus, we also present a method to combine both CET and cryo-EM data, in which the asymmetric reconstruction from CET data cubes are used as an initial model for searching the absolute orientations for cryo-EM images, thus resulting a high-resolution asymmetric reconstruction of cryo-EM data that preserves the asymmetric region. Similarly, in Chapter 2 we demonstrate that our classification and reconstruction algorithm might be superior to some traditional classification and reconstruction methods developed for cryo-EM when applied to CET data cubes since results from our algorithm is free from reference bias, but our method is much more computationally expensive than the traditional ones. Thus, in analyzing tomograms in Chapter 3, reconstructions are obtained from a subset of the available data using our method to guarantee no initial reference bias, then the reconstructions are used in a correlation-based template matching algorithm, which is used widely in processing cryo-EM data, to analyze all available tomograms in order to reduce computation. Finally, the physical model developed in Chapter 6 is also useful in structural determination from experimental data. Stable intermediate structures can be predicted

from such a model and these structures can be used as references to sort a set of heterogeneous cryo-EM images into more homogeneous subsets. In Ref. [14], Brink et. al., are able to obtain higher resolution structures of Fatty Acid Synthase (FAS) by using intermediate states predicted from such a model in a multi-reference refinement reconstruction algorithm.

The hybrid approaches we discussed above are a straight forward way to incorporate different imaging data sets. Perhaps a more ambitious and interesting extension is to incorporate structural information from a wide-range of experiments such as mass-spectrometry, solution x-ray scattering, FRET spectroscopy etc, which is a nontrivial task. The model-based statistical approach we presented here might serve as a foundation for incorporating all of these experimental information in an integrative way that will shed more light on the structural behavior of super-molecular complexes than is possible from any single experimental approach alone. A recent success in studying the nuclear pore complex using such an integrative approach both confirms the possibility and reiterates the importance of such a comprehensive approach to structural biology [4].

## CHAPTER 2

### A MAXIMUM-LIKELIHOOD APPROACH TO THE CLASSIFICATION AND ALIGNMENT OF CET CUBES

#### 2.1 Introduction

Cryo-Electron Tomography (CET) is an imaging technique that is able, via transmission electron microscopy plus computation, to determine the 3-D structure of one-of-a-kind biological specimens up to the size of a small whole cell. When applied to a cell, CET is an *in situ* technique, i.e., it provides images of macromolecular assemblies within the cell in a hydrated state that is close to their native environment and with their natural spatial relationship to other macromolecular assemblies [76, 46]. However, CET has limitations. In order to avoid destruction of the specimen, the total electron dose is minimized, which results in SNR for the original images that is typically much lower than 0.1. Furthermore, for the following two reasons, the angle coverage for the projection data is limited. First, the tilt range of the goniometer which holds the specimen is limited to about  $\pm 60^\circ$  for mechanical reasons [116]. Second, the thickness of the specimen is itself a limit. Perpendicular to the sample grid, the specimen is typically  $0.5\mu\text{m}$  thick. However, as the sample grid is tilted, the distance through which the electron beam must penetrate increases. For example, the penetrating depth at a tilt of  $60^\circ$  is twice that at  $0^\circ$  tilt, and the penetration depth at  $0^\circ$  is already near the maximum achievable penetration depth. Thus, the projection images one can acquire for CET are typically at a tilt range of  $\pm 60^\circ$ . By the Projection-Slice Theorem, the limited range of angles for the projection data leads to a missing wedge of data in 3-D reciprocal space which results in ge-

ometric distortion in the direction parallel to the tilt axis [74]. The challenges of low SNR and limited angle coverage for the projection data lead to a noisy 3-D tomographic reconstruction problem that has a missing wedge in reciprocal space. The resulting resolution achieved in the 3-D tomogram of CET is on the order of 10nm while reconstructions based on single cryo-EM images of multiple identical objects can achieve resolutions less than 1nm.

In a typical infected cell there are tens to hundreds of virus particles at various stages of maturation. Therefore, if the stages can be classified and information from the particles at the same stage can be combined, both signal to noise and resolution improvement can be improved. There are two fundamental challenges in combining information from different particles. First, the different particles are differently oriented and so the subcubes of the tomogram must be oriented before the information can be combined. Second, different particles come from different classes (e.g., maturation stages) and each subcube of the tomogram must first be classified before the information it contains can be combined with the information from other subcubes and classification is difficult due to the low SNR, the low resolution, and the missing wedge of the data. Particle orientation and particle classification into small numbers of homogeneous classes has been used previously in single-particle cryo-EM [33, 119, 98]. However, the data in single-particle cryo-EM and CET are quite different: The data from the CET experiment are 3-D subcubes from the tomographic reconstruction, including distortions due to problems such as the missing wedge of data, while the data in single-particle cryo-EM are 2-D projection images. These differences are the origin of two challenges. First, the 3-D nature of CET data increases the computational burden. Second, the CET data is imperfect since it is itself the result of a reconstruction algorithm and most importantly has a

missing wedge in reciprocal space, which hampers both the alignment and classification of particles using methods developed for single particle analysis [31].

A standard approach [33] is to classify information into homogeneous classes, possibly even including orientational classes, and then reconstruct. In this type of approach, standard cross-correlation, which is used extensively in the alignment of single-particle analysis, has been shown to be not accurate in aligning data cubes with missing wedges, especially for macromolecules with high symmetry such as many viruses, which often possess icosahedral symmetry [100]. To compensate for the missing wedge, alignment algorithms based on modified cross-correlation functions that consider the non-zero overlapping regions in reciprocal space of the two data cubes have been proposed by various researchers [8, 31, 100, 116]. Regardless of what similarity measure these algorithms use, one common feature is that the data cubes are rotated over all possible ranges to search for the correct pairwise alignment. Rotations will inevitably lead to interpolation of the original data cube due to the discrete nature of the data. Since the typical SNR of these tomograms is much less than 1, any form of interpolation will give rise to significant errors to the rotated data cube and hence affect alignment accuracy. Thus, pairwise alignment of noisy data cubes may not be the optimal approach. The maximum likelihood approach described here uses a mathematical model that describes the virus in continuous coordinates and rotates the model in continuous coordinates before sampling and comparing with data cubes thereby avoiding the need for interpolation.

A more fundamental issue than the issue of interpolation is that the alignment and classification depend on the accuracy of one another. Accurate alignment requires that the subset of data to be relatively homogeneous, while ac-

curate classification requires that each data subcube be accurately aligned, so that the differences between classes are due to structural differences rather than orientational differences. One approach to the interdependency of alignment and classification is to use iterative refinement to progressively get better classifications and alignments in alternating steps (e.g., Ref. [8]). Alternatively, the alignment and classification can be performed jointly (e.g., as described in the conclusion of Ref. [31]).

In this chapter, we describe an approach that goes beyond joint alignment and classification to joint alignment, classification, and restoration (SNR and resolution improvement) by maximum likelihood (ML) estimation. Joint ML classification, alignment, and reconstruction has been used extensively in single-particle cryo-EM [73, 98], but not in CET. Joint ML for CET has the same desirable properties as joint ML for single-particle cryo-EM, e.g., statistical hedging of uncertainty in class and orientation which is important for these low SNR, low resolution 3-D data cubes. For tomograms of purified groEL/groES, a maximum likelihood approach to classification and reconstruction has been described [99] which uses a mathematical formulation in which the missing reciprocal space data is treated as nuisance parameters in an expectation maximization algorithm, which is different from the formulation in this chapter, and which has not been applied to *in vivo* whole-cell tomograms.

## 2.2 Mathematical model for the object

The electron scattering intensity of a 3-D object, denoted by  $\rho(\mathbf{x})$ , can be represented as a linear combination of basis functions, denoted by  $\varphi_i(\mathbf{x})$ , where each

basis function is a continuous function of the three spatial coordinates (denoted by  $\mathbf{x} \in \mathbb{R}^3$ ), i.e.,

$$\rho(\mathbf{x}) = \sum_i d_i \varphi_i(\mathbf{x}) \quad (2.1)$$

where the unknown weights which determine the 3-D structure of the virus are the  $d_i$ . The selection of basis functions can be tailored specific to the macromolecular complex of interest.

Sulfolobus Turreted Icosahedral Virus (STIV) is a spherical virus with icosahedral symmetry. Because the icosahedral group has 60 rotation symmetry operators and the particles are spherical, it is natural to represent such an object by a spherical harmonics series using icosahedral harmonics, but other representations of the complete particles (e.g., voxels) are also possible. For any representation as a linear combination of basis functions, the statistical model and ML estimation approach of Section 2.3 can be applied.

In this chapter, the 3-D scattering intensity of a complete STIV particle is described by

$$\rho(\mathbf{x}) = \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} h_{l,p}(|\mathbf{x}|) \Psi_{l,m}(\theta(\mathbf{x}), \phi(\mathbf{x})), \quad (2.2)$$

where the unknown weights  $d_i$  of Eq 2.1 now require triple indices, i.e.,  $d_{l,m,p}$ . One advantage of this model is that rotational symmetry can be built into the model by simply restricting the functions  $\Psi_{l,m}$  to be a basis for the rotationally symmetric subspace of functions on the sphere, i.e., to the rotationally symmetric subspace of the space spanned by spherical harmonics. This causes a reduction in the number of parameters which need to be estimated. In the case of icosahedral symmetry, the reduction is by a factor of 60. This reduction both reduces the computational cost (Section 2.4) and improves the quality of the estimates. Concretely, this is achieved by defining  $\Psi_{l,m}(\theta, \phi)$  to be a linear com-



combination of spherical harmonics  $Y_{l,m'}(\theta, \phi)$  [53, Eq. 3.53] ( $m' \in \{-l, \dots, +l\}$ ) chosen so that the collection of  $\Psi_{l,m}(\cdot, \cdot)$  span the rotationally symmetric subspace and so that  $\Psi_{l,m}(\cdot, \cdot) \in \mathbb{R}$  so that  $\rho(\cdot) \in \mathbb{R}$  for any choice of  $d_{l,m,p} \in \mathbb{R}$ . The radial basis functions, denoted by  $h_{l,p}(\cdot)$  and derived by Sturm-Liouville theory [21, Ch.7], are linear combinations of spherical Bessel functions [53, Eq. 16.9] which satisfy  $h_{l,p}(r) = 0$  for  $0 \leq r \leq r_1$  and  $r_2 \leq r$ . The possibility of  $r_1 = 0$  is permitted and the possibility of  $r_1 = 0$  and  $h_{l,p}(0) \neq 0$  for some values of  $l$  and  $p$  is also permitted. The 3-D Fourier transform of  $\rho(\mathbf{x})$ , denoted by  $P(\mathbf{k})$ , is

$$P(\mathbf{k}) = \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \Psi_{l,m}(\theta(\mathbf{k}), \phi(\mathbf{k})) \quad (2.3)$$

where  $H_{l,p}(\cdot)$  is the spherical Hankel transform of  $h_{l,p}(\cdot)$ , which can be computed analytically because of the choice of  $h_{l,p}(\cdot)$ .

### 2.3 3-D cube formation model

Let  $\Sigma(\mathbf{k})$  denote the 3-D Fourier transform of a noise free (i.e., ideal) subcube extracted from the CET 3-D reconstruction. Then,

$$\Sigma(\mathbf{k}) = G(\mathbf{k}) W(\mathbf{k}) \exp(-i2\pi \mathbf{k}^T \mathbf{x}_0) P(R(\alpha, \beta, \gamma)^{-1} \mathbf{k}) \quad (2.4)$$

where  $\mathbf{x}_0$  describes the translation between the center of the coordinate system and the center of the object,  $(\alpha, \beta, \gamma)$  are Euler angles describing the rotation of the object,  $W(\cdot)$  is the 3-D frequency response describing the loss of the missing wedge of data, and  $G(\cdot)$  is the 3-D CTF for the CET measurement and reconstruction process. Please note that  $P(\cdot)$  depends linearly on the unknown coefficients  $d_{l,m,p}$ .

According to the Projection Slice Theorem, the 2-D projection of a 3-D function in a given projection direction is, after transformation to reciprocal space, a slice (i.e., 2-D plane) through the origin of the reciprocal space transformation of the 3-D function, where the plane is perpendicular to the projection direction. Thus, assuming reconstruction using a perfect interpolation filter, the missing data in the reconstruction for a single tilt experiment of  $\pm\theta$  will be the region sandwiched between the two planes that form angles of  $\pm\theta$  with the tilt axis. This region is shaped like a wedge, thus the term “missing wedge”. The function  $W(\mathbf{k})$  in Eq. 2.4 accounts for the missing wedge by being 1 where the reciprocal space data is not missing and 0 where the data is missing, i.e., a binary mask. Let  $\mathbf{n}_1$  and  $\mathbf{n}_2$  be unit vectors normal to the two planes that are the boundaries of the missing wedge. Then

$$W(\mathbf{k}) = \begin{cases} 0, & (\mathbf{k} \cdot \mathbf{n}_1 > 1) \cap (\mathbf{k} \cdot \mathbf{n}_2 > 1) \\ 0, & (\mathbf{k} \cdot \mathbf{n}_1 < -1) \cap (\mathbf{k} \cdot \mathbf{n}_2 < -1) \\ 1, & \text{otherwise} \end{cases} \quad (2.5)$$

where  $\cap$  indicates the logical “and” operation.

Assuming that the virus particles are roughly randomly oriented relative to the electron beam of the microscope and are sufficiently numerous, it is expected that no region of reciprocal space will be in the missing wedge of all the particles. Therefore, by combining information from different particles, the method described in this chapter can greatly reduce the effect of the missing wedge.

In the current software, the 3-D CTF for the CET measurement and reconstruction process, introduced in Eq. 2.4 and denoted by  $G(\mathbf{k})$ , is assumed to be 1 at all spatial frequencies. Therefore, the work described in this chapter accounts for the low SNR, the low resolution, and the missing wedge of the CET result

but not other limitations of the CET result. Such limitations, if linear, could be incorporated in  $G(\mathbf{k})$ .

As discussed previously, the CET cube has quite low SNR. The 3-D cube noise is assumed to be an additive white zero-mean known-variance Gaussian noise where the variance is actually estimated in a preliminary calculation. The assumption of additive Gaussian noise is motivated by the simplicity of the solution of the estimation problem formulated in Section 2.4.1.

Because  $P$  in Eq. 2.4 is a linear function of the unknown coefficients  $d_{l,m,p}$ , Eq. 2.4 coupled with the noise assumptions of the previous paragraph can be rewritten as follows. Let the pixels of a reciprocal space subcube be arrayed in a vector denoted by  $y$  and let the voxel noise be similarly arrayed in a vector denoted by  $v$ . Let the unknown coefficients  $d_{l,m,p}$  be arrayed in a vector  $d$ . Let  $z = (\mathbf{x}_0, \alpha, \beta, \gamma)$ . Then Eq. 2.4 and the noise assumptions of the preceding paragraph imply that there exists a matrix, denoted by  $L$  which depends on  $z$ , such that the relationship between  $y$ ,  $d$ , and  $v$  is

$$y = L(z)d + v. \quad (2.6)$$

The elements of  $L$  are defined by Eq. 2.4 and are products of icosahedral harmonics and radial basis functions. When multiple classes are considered, there are multiple  $d$  vectors, one for each class, and potentially multiple  $L$  matrices if, for example, different classes have different symmetry assumptions and therefore different dimensions for  $d$  even though they achieve the same resolution.

## 2.4 Statistical estimation

### 2.4.1 Maximum likelihood estimation

Eq. 2.6 implies a likelihood function for estimating  $d$  from  $y$  by maximum likelihood estimation, in particular, the likelihood function conditional on  $z$  is  $p(y|d, z) = \mathcal{N}(L(z)d, \Sigma)$  where  $\Sigma$  is the covariance of  $v$  and  $\mathcal{N}(m, V)$  is the multi-variable Gaussian probability density function (pdf) with mean vector  $m$  and covariance matrix  $V$ . Then the unconditional likelihood function (which can be computed once the pdf on  $z$ , denoted by  $p(z)$ , is specified) is

$$p(y|d) = \int_z p(y|d, z)p(z)dz. \quad (2.7)$$

The pdf for  $z$  which is used in Section 2.6 is uniform over a 3-D sphere in  $\mathbf{x}_0$  and uniform over all orientations in  $(\alpha, \beta, \gamma)$  (i.e., Haar measure on  $\text{SO}_3$ ). The subcube extracted from the tomogram is a cube to which a spherical mask is applied in order to remove objects adjacent to the virus particle. Before masking,  $\Sigma$  is proportional to the identity, but masking introduces correlation. However, because the mask is large, the correlation is mostly between nearest-neighbor pixels and is not accounted for in our current software.

Given the likelihood function (Eq. 2.7), the goal is to estimate  $d^\eta$  for each class where class is indexed by  $\eta$ . In order to simplify notation, only the equations for a one-class problem are written so the index  $\eta$  is not necessary. For a multi-class problem with  $N_\eta$  number of classes, each iteration of the EM algorithm involves computing the same integrals (Eqs. 2.13 and 2.14) and solving the same system of linear equations (Eq. 2.12) as was done in the one-class problem but now for each of  $N_\eta$  different classes. In the maximum-likelihood sense, the estimate of

$d$ , denoted by  $\hat{d}$ , is the solution that maximizes the likelihood function (Eq. 2.7), i.e.,

$$\hat{d} = \arg \max_d p(y|d). \quad (2.8)$$

Using the same type of approach as was employed in Ref. [24], an expectation maximization (EM) algorithm is derived for computing the maximum likelihood estimate for  $d$ . An EM algorithm is an iterative algorithm that progressively finds a better estimate of the model parameter  $d$  at each iteration. Specifically, at iteration  $n$  of the EM algorithm, a quantity  $Q$  is computed based on the estimate of the model parameter from the previous step, denoted by  $d_{n-1}$ , where  $Q$  is defined by

$$Q(d|d_{n-1}, y) = \int_z (d^T L^T L d - 2y^T L d) p(z|d_{n-1}, y) dz \quad (2.9)$$

$$= d \left\{ \int_z L^T L z p(z|d_{n-1}, y) dz \right\} d - 2 \left\{ \int_z y^T L(z) p(z|d_{n-1}, y) dz \right\} d \quad (2.10)$$

The second equality is due to the linearity of the integral and the fact that  $d$  does not depend on  $z$ . Then the estimate of the model parameter at the  $n$ th step, denoted by  $d_n$ , is

$$d_n = \arg \max_d Q(d|d_{n-1}, y). \quad (2.11)$$

Notice that Eq. 2.10 is a quadratic form in  $d$ , for which the solution that maximizes the quantity  $Q$  can be obtained by solving the linear system

$$g = F d_n \quad (2.12)$$

where

$$F = \int_z L^T(z) L(z) p(z|d_{n-1}, y) dz \quad (2.13)$$

$$g = \int_z L^T(z) y p(z|d_{n-1}, y) dz. \quad (2.14)$$

### 2.4.2 Practical issues

Similar to most cryo EM reconstruction algorithms, the resolution of the whole-particle reconstruction is increased in a series of steps. Resolution of the model is controlled by truncating the  $l$  and  $p$  sums in Eq. 2.2 to  $0 \leq l \leq l_{\max}$  and  $1 \leq p \leq p_{\max}$ . Five reconstructions of progressively improved resolution, denoted Step 0 to Step 4, are calculated, in which the  $l_{\max}$  and  $p_{\max}$  values for each step are the same as were used in Ref. [119].

Step 0 is estimation of a spherically symmetric model (i.e., only the  $l = 0$   $d_{l,m,p}$  coefficients in Eq. 2.2 are used) based on the available data. In this case, no knowledge of the particle orientation for each data cube is needed and therefore a one-class reconstruction is essentially a linear least squares problem which is solved by standard methods. Steps 1–4 use models with icosahedral but not spherical symmetry so the expectation-maximization (EM) algorithm of Section 2.4.1 is used. EM is an iterative algorithm that requires initialization. Since the EM algorithm converges to a local maximum of the likelihood function, multiple initializations are needed to ensure that the algorithm converges to the global maximum. One of the initial conditions is the answer from the previous step augmented with zeros for those  $d_{l,m,p}$  coefficients that did not occur in the previous step. The remainder of the initial conditions are Gaussian pseudo-random perturbations of the initial condition derived from the answer from the previous step. The parameters that determine this process are identical to those in Ref. [119]. The final answer for a step is obtained by selecting the solution that has the highest likelihood value among all of the initializations.

For the whole-particle reconstructions the nuisance parameters are the class membership label of the particle and the orientation of the particle. Integration

over the orientational nuisance parameters is done by the 5000 abscissa integration rule of Ref. [119].

## 2.5 Synthetic data from the x-ray crystallographic structure of Hong Kong 97

In order to study the performance of the proposed algorithm, synthetic data based on the x-ray crystallographic structure of the capsid of the bacteriophage Hong Kong 97 (HK97) are generated and processed. During its maturation cycle, the capsid of HK97 undergoes extensive conformational changes from a stable procapsid structure of approximately 450 in diameter to a final infectious particle structure of approximately 650 in diameter which is able to withstand over 50 atm of internal pressure due to genome packaging [38]. The atomic structures of both procapsid and mature capsid of HK97 are known from x-ray crystallographic studies, thus making it an excellent model to assess the performance of the proposed algorithm in the application of studying virus maturation *in situ*.

A set of 3-D cubes each containing one HK97 virion with the distortions introduced by noise, limited tilt angle, and reconstruction by IMOD [65, 80] is computed. The set is a mixture of cubes where each cube shows either the procapsid or the capsid. Each individual cube is computed by the following algorithm:

1. Pick a pseudo-random number to determine whether the cube will show the procapsid versus the capsid, which are both shown in Figure 2.5. The

- probability that the 3-D cube shows the procapsid (capsid) is 0.6 (0.4).
2. Compute a 3-D density map of the particle chosen in the previous step (procapsid or capsid) with a 1.6nm sampling rate from the x-ray crystallographic structure. In this density map, the particle is in the standard orientation.
  3. Rotate the 3-D density map by applying the 3-D rotation described by one of the 5000 abscissas used to integrate over orientations where the choice is uniformly distributed over the 5000 possibilities.
  4. Generate projection images based on the rotated model. The range of tilt angles is from  $-\theta$  to  $+\theta$  and the tilt angle increment is  $2^\circ$ . (The value of  $\theta$  is specified below).
  5. Add zero-mean white Gaussian noise with a variance that is determined in order to achieve the desired SNR (specified below) to each projection image.
  6. Compute a 3-D reconstruction of the virus particle using IMOD [65, 80].

This set of cubes showing a mixture of procapsid and capsid is then the input to the algorithms described in this chapter. Two different synthetic experiments are conducted, which are described in Sections 2.5.1 and 2.5.2.

### **2.5.1 Performance as a function of SNR and tilt range microscopy parameters**

To test the proposed algorithm under different experimental conditions, synthetic CET cubes are computed by the above procedure for a range of SNR val-



ues, specifically, 0.1, 0.05, 0.01, and 0.005, and a range of single-tilt tilt values, specifically tilts every  $2^\circ$  in the range  $\pm\theta$  where the value of  $\theta$  is  $70^\circ$ ,  $60^\circ$ , or  $50^\circ$ . For each SNR and tilt range, sixty data cubes of  $51^3$  voxels are generated. Simultaneous reconstruction and classification assuming two classes of particles are performed based on the data cubes from each experimental setting.

Cross-sections of the simulated data cubes, one for each experimental setting, are shown in Figure 2.6. Two of the SNR values, 0.01 and 0.005, are quite challenging, note that the signals are barely visible in the corresponding images of Figure 2.6, and the purpose of using these low SNR values is to demonstrate the ability of the algorithm to perform at low SNR. Figure 2.7 shows slices through the origin in the  $x$ - $z$  plane of simulated data cubes in reciprocal space for tilt range of  $\pm 70^\circ$ ,  $\pm 60^\circ$ , and  $\pm 50^\circ$ , which illustrate the characteristic wedge shape of the missing Fourier data in the tomographic reconstruction of CET. Since the tilt axis in these experiments is the  $y$ -axis, the missing wedge appears as a pair of triangles in the  $x$ - $z$  plane. As the tilt range is reduced, the missing wedge increases which indicates the loss of more Fourier data.

One way to assess the performance of the proposed algorithm is to measure the agreement between the reconstruction and the 1.6nm reference structure at different spatial frequencies using Fourier Shell Correlation (FSC). The standard FSC criterion is used: two structures agree at a specific spatial frequency if the correlation value is greater than 0.5 for all frequencies less than the specific frequency and the inverse of the lowest spatial frequency at which the correlation is less than 0.5 is defined to be the resolution of the reconstruction. Notice that this resolution measure describes the similarity of the reconstruction and the reference structure and is not the resolution measure used in single particle cryo

EM studies which describes the quality of the set of projection images.

As mentioned previously, no explicit orientations and class labels are assigned for each data cube. Instead the algorithm estimates a model for each class that best fits the entire data set. However, after the estimated model for each class is obtained, an estimate of the orientation for each data cube, denoted by  $\hat{\theta}$ , can be calculated by

$$\hat{\theta} = \arg \max_{\theta} p(y|\hat{d}, \theta) \quad (2.15)$$

where  $\hat{d}$  denotes the estimate of the model parameters. Similarly, a class label for each data cube can also be computed. This leads to a second way to measure the performance of the algorithm based on the accuracy of the class label and orientation for each data cube. When each synthetic data cube was computed, the orientation of the particle was selected from the set of abscissas used in the expectation step of the expectation maximization algorithm and the  $\arg \max_{\theta}$  in Eq. 2.15 is also computed over the same set. Thus, it is possible for the estimated orientation to be exactly that of the true orientation, which eliminates the effect of discretization error in searching for the correct particle orientation.

Table 2.1 summarizes the resolution results for the first experiment. For each entry in the table, two resolutions are reported, which correspond to the reconstructions of the procapsid and capsid respectively. The resolution of the reconstructions for various experimental conditions are between 3.27nm to 4.3nm, while the Nyquist limit is 3.2nm, indicating that they are in good agreement with the reference structures. This analysis shows the algorithm is able to successfully classify and reconstruct the two underlining structures reliably from noisy data cubes that are further degraded by the missing wedge effect under various experimental conditions. Even when the SNR value is 0.005, the al-

gorithm is still able to obtain reconstructions that have similar quality to that obtained based on data with much higher SNR. This illustrates the robustness of the maximum-likelihood approach against noise.

The resolution of the procapsid reconstruction is always slightly better than that of the capsid reconstruction. Several factors might contribute to this phenomenon. First, the procapsid is more common in the mixture of data cubes (60% procapsids versus 40% capsids). Furthermore, the procapsid occupies a slightly smaller cube in the  $51^3$ -voxel data cube than the capsid, which results in slightly finer sampling in reciprocal space due to the additional zero padding, thus increasing its correlation values at every spatial frequency in the Fourier Shell Correlation assuming the two structures are similar. A third factor is that the capsid has a smoother surface than the procapsid and for that reason it may be that the orientation estimates for the procapsid are more accurate than for the capsid leading to higher resolution in the reconstruction of the procapsid.

Tables 2.2 and 2.3 summarize the orientation estimation results for the first experiment. The number of data cubes for which the algorithm did not estimate the orientation correctly are reported in Table 2.2. For the misaligned data cubes, the average error is quantified in Table 2.3 as a function of the range of tilt angles used, i.e., as a function of the size of the missing wedge. The number of misaligned data cubes increases as SNR decreases, reflecting the fact that as noise degrades the signal, it is harder to detect orientation reliably. The number of misaligned data cubes also increases as the tilt range gets smaller. This is also expected, since as the amount of missing data increases, the information needed to correctly estimate the particle orientation might not be available.

While performance degrades as SNR decreases, note that even with the sig-

nificant misalignment at SNR=0.005 (13%, 17%, 27% for tilt range  $\pm 70^\circ$ ,  $\pm 60^\circ$ ,  $\pm 50^\circ$ , respectively), the algorithm is still able to obtain a reasonable quality of reconstruction for both the procapsid and the capsid in terms of resolution. This is due to two factors. First, the misalignment is not great as is shown in Table 2.3. Second, at each iteration, the algorithm effectively considers all possible orientations that the particle could assume in each data cube and weights their importance according to their similarity to the structure estimated at the previous iteration. Even though the most probable orientation is not the correct orientation for a particular data cube due to the noise corrupting the data, the algorithm computes a structure that still takes the true orientation of the particle into account, thus the final structure is still in good agreement with the reference structure. In essence, the ML approach reflects a soft alignment and classification among the data set rather than determining ahead of time what orientation and class each data cube belongs to. This might be the reason that the ML approach is less sensitive to initial condition, unlike some of the other classification methods. Also, this soft alignment and classification strategy might be one factor that makes the ML approach to be more robust in much lower SNR environment such as 0.01 and 0.005.

### **2.5.2 Performance as a function of number of data cubes**

In the second synthetic experiment, subsets of 10, 20, and 30 cubes from the set of 60 data cubes simulated at SNR=0.01 and a tilt range of  $\pm 60^\circ$  are used for reconstructions. The goal of this experiment is to evaluate how the performance of the algorithm improves as the number of data cubes used increases.

For this experiment, the resolution of the reconstructions based on different numbers of data cubes are shown in Table 2.4. Even using ten data cubes, the resolution at which the reconstruction and reference agrees is 3.45nm and 4.36nm for the procapsid and capsid respectively. For the reconstruction of the capsid, the resolution improves progressively from 4.36nm to 3.73nm as the number of cubes used in the reconstruction increases. However, for the procapsid reconstruction, the resolution using only 10 cubes is much better than the resolution achieved for the capsid with 10 cubes and the resolution does not improve until 60 data cubes are used. To assess the quality of the reconstructions in more detail, FSC curves between the reconstruction and reference structures for the procapsid and capsid are shown in Figure 2.8(a) and (b), respectively. Comparing the FSCs for different reconstructions, the overall curves progressively shift to the upper right as the number of data cubes used in the reconstruction increases, which is an indication that the quality of the reconstructions at every spatial scale improve as the number of cubes used increases. Therefore, the fact that resolution does not change for the procapsid until 60 data cubes are used is due to the threshold value applied to the FSC curves rather than due to the algorithm. This synthetic example illustrates the asymptotic property of the maximum likelihood estimator, in which the estimator converges to the exact truth as the number of observations increases, no matter how noisy the data is.

To illustrate the resolution improvement achieved by combining data cubes, FSC curves between the simulated data cubes and the reference structures are also computed and displayed in Figure 2.8. Averaging the 40 (20) resolutions of the 40 (20) capsids (procapsids) gives the value of 11.72nm (11.84nm), which is much lower than the resolution tabulated in Table 2.4. Among the factors contributing to the resolution gain achieved by the algorithm of this chapter

are the following. First, by combining different data cubes, the SNR is significantly improved. Secondly, the missing wedges in the original data cubes lead to significant blurring and geometric distortion in the original data cubes. The algorithm of this chapter fills in the missing wedges because it combines data from cubes with different orientations and therefore different positions of the missing wedge, thus alleviating the anisotropic effect result from the missing wedge.

The alignment errors and classification errors are summarized in Table 2.4. Using even as few as 10 data cubes, the algorithm is able to separate the pro-capsid and capsid structures reliably. For orientation estimation, the algorithm is not able to correctly estimate the orientations of a few data cubes in each case. Since the smaller set of data cubes is always a subset of the bigger set, it is possible to examine how the algorithm performs at the misaligned data cubes with a bigger data set. Interestingly, those that are not aligned correctly in the smaller set are aligned correctly in the bigger sets. This is expected, since as the resolution of the reconstruction improves, the algorithm is able to get a correct orientation estimation that it is not able to obtain with a lower quality reconstruction.

One final remark is that the classification is always reliable in the synthetic experiment. This might be partly due to the fact that the two structures used in generating the synthetic data, although biologically related, might be sufficiently different in a structural sense such that the algorithm is always able to distinguish between them. However, what is more important, is that since the classification, alignment, and reconstruction are performed simultaneously, no misclassification is done by grouping particles with similar missing wedge

orientations, which is observed in some of the other iterative classification and alignment processes.

## 2.6 Experimental data from STIV infected *Sulfolobus sulfataricus* cells

Whole-cell tomograms of STIV infected *Sulfolobus sulfataricus* cells were collected as is described in Ref. [35]. The ML method is employed on four different tomograms in order to analyze viral particles assembly in different cells. Representative sections are shown in Figure 2.1. A typical viral-infected *Sulfolobus* displays multiple pyramid-like protrusions on the cell surface that cause dramatic alteration of cell morphology (Fig 2.1a). Viral particles with either full or empty cores were observed, representing DNA-filled virions and DNA-free procapsids. Virions organizing as quasi-crystalline arrays have been observed, distributing distinctly from procapsids that mostly scatter outside the arrays [35]. Some ruptured cells gushing out cytoplasmic materials were captured (Fig 2.1b, c). The breakage site featured the overhang of pyramid fragments reiterates that the pyramid structures are more fragile than normal cell wall, which is protected with hexagonally arranged surface protein layer (S-layer). Intrapyramidal bodies have been observed to associate with pyramids. Even though electron-dense bodies also present in the uninfected cells, the compositions and functions of those bodies remain to be discovered.

93, 68, 62, or 141 CET cubes, each containing one virus particle, are extracted from a cell tomogram by cross-correlation with a simple spherical template. The missing wedge causes blurring of the outline of the virus capsid in the  $z$  direc-

tion which can be seen in  $x$ - $z$  and  $y$ - $z$  cross sections of virus particles (not shown).

By manual inspection, virus particles in the cell tomogram are either packed with genome or are empty. This is not surprising because packaging kinetics can be fast (in dsDNA bacteriophages, the time constant of the packaging kinetics are on the order of couple minutes [36, 101, 129] ), in which case only a small fraction of particles will be in a partially-packed state. Hence, the reconstruction calculation is based on the assumption that two classes of particles exist. Furthermore, the STIV particle possess icosahedral symmetry, so such a symmetry constraint is also enforced in the model.

One measure of the quality of the reconstructions is the quality of the orientations computed from the maximum likelihood estimates. One measure of the quality is to average the oriented particles and check whether structures such as turrets are preserved in the averaging. As is shown in Figure S3c,d of Ref. [35], the turrets are preserved.

3-D visualizations of the full and empty reconstructions from two of the four tomograms are shown in Figure 2.2. There is a clear distinction between the reconstructions for Class 1 and 2, in particular, the Class 2 reconstruction has reduced central density and therefore represents the empty virus capsid. Thus, the algorithm is able to separate the full and empty viruses in the cell tomogram. Furthermore, the icosahedral shape of the capsid and the turrets at the five-fold axes of the capsid, both features that are severely distorted in the CET cube due to the missing wedge to the degree that they are not visible in the  $x$ - $z$  and  $y$ - $z$  cross-sections of the CET data subcubes, are clearly present in both reconstructions, providing evidence that by combining multiple noisy 3-D subcubes with different orientations, the missing data in reciprocal space is successfully recov-



ered.

Each of the four tomograms was processed separately. Therefore there are four independent reconstructions of the full and empty particles. In order to demonstrate that the maximum likelihood estimator chooses the same class definitions in each tomogram, the Fourier Shell Correlation (FSC) was computed for all pairs of full and for all pairs of empty reconstructions with the results shown in Figure 2.3. Using an FSC cutoff of 0.5, the full (empty) particle reconstructions agree to 6.61 nm (6.67 nm) while the resolutions of the individual reconstructions (computed by FSC between even and odd numbered data sub-cubes) are 6.1 nm (6.9 nm). Therefore the maximum likelihood estimator, which determines its own classes once the number of classes has been set by the user, is determining the same classes for each tomogram.

For the calculations on STIV infected *Sulfolobus sulfataricus* cells described in Ref. [35] and in the sectional images of Figure 2.1, the particles tend to cluster within the cell and the clusters tend to approximately have the structure of a lattice. In Supplemental Material Section A.1, algorithms are described for determining the lattice constants and the orientational heterogeneity of the particles occupying the lattice.

## 2.7 Discussion and conclusion

The CET studies of STIV infection in intact cells revealed valuable insights of assembly and maturation of STIV and inner-membrane containing viruses in general and uncovered the ultra-structural alterations with fine details. Combining with computational analysis, structures like transiently populated pro-

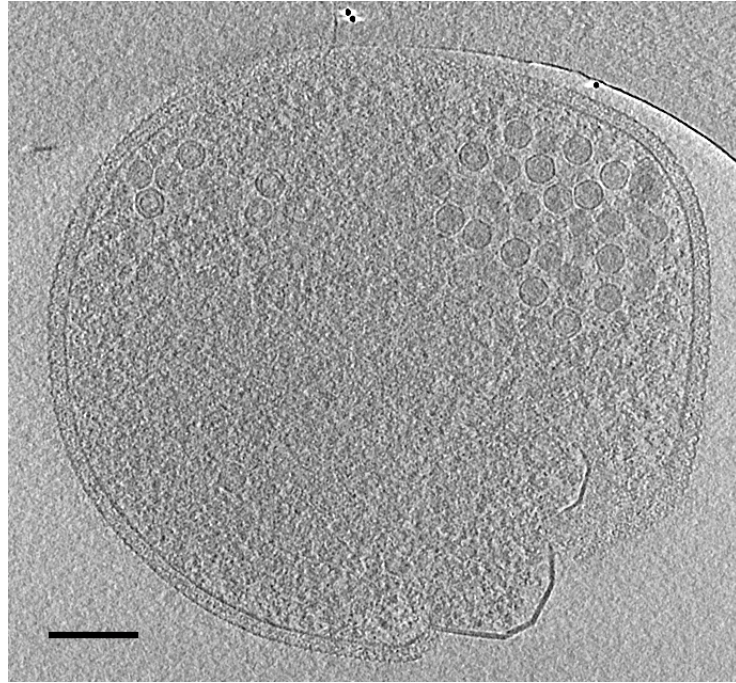
capsids and unstable assembly intermediates can be visualized in their original context and characterized at low-nm resolution without the need of purification or concerns of artifacts from sample.

The algorithm proposed in this chapter reflects a different approach than those in Refs. [8, 31, 100], in that no explicit alignment and classification are performed on objects in the data set. Instead, a complete statistical model is described for each class of particles among the available data, then model parameters are obtained by maximizing the log-likelihood function. The specific log-likelihood function used in the statistical model is essentially the dissimilar function used in Ref. [8] and with normalization is also the constrained cross-correlation used in Ref. [31], all of which explicitly account for the missing wedge of the data. Thus, the proposed algorithm is in accordance with the processing methods proposed elsewhere, in that all the algorithms seek to optimize the same cost function with respect to the data set, just using different approaches.

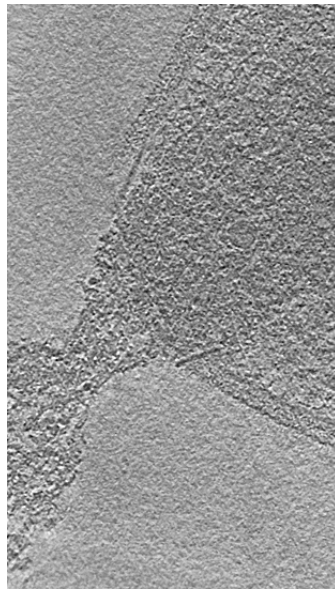
The proposed algorithm uses a soft alignment and classification strategy rather than the hard alignment and classification strategies used elsewhere, in that each class average is computed by averaging weighted sum of all data cubes at all possible orientations in each step of the maximization process. The weight is proportional to the similarity of the rotated data cube and the class average from the previous iteration. As the class average improves, the averaging will be weight mostly toward the data cubes within that class and the correct orientation of each data cube. Thus, as the algorithm converges, the underlying structures will be obtained. This alignment and classification strategy might be important at the initial stage, when the class averages are poor. In this

situation an algorithm can lock onto grouping data cubes that contain particles with similar missing wedge orientations rather than similar structures. However, with the soft classification, the class average might not have a strong bias in this situation, making it easier to escape such a local minimum.

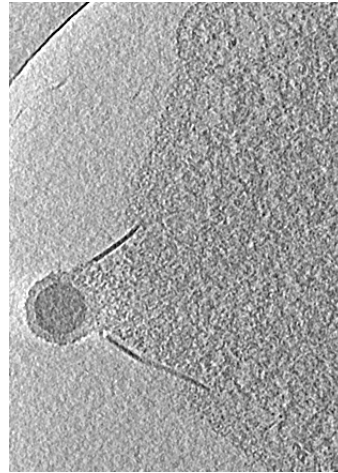
Whole-cell Cryo-Electron Tomography is a unique imaging technique, in that it allows the visualization of macromolecules in their native cellular environment. However, due to the technical limitations, it has inherent missing data in reciprocal space, and the missing data implies that conventional techniques to not reliably process the resulting 3-D real-space cubes. The maximum likelihood (ML) approach, which describes a complete statistical model of the tomographic data that explicitly accounts for the missing wedge effect, can reliably classify heterogeneous particles in the cell tomogram and accurately align each to a common coordinate system, so that a higher resolution structure that is free of the missing wedge effects can be obtained. Furthermore, the ML approach is robust against noise, in particular, in simulation, it performs relatively well even at an SNR of 0.005. As the Whole-cell Cryo-Electron Tomography technology matures, better cameras will allow the acquisition of projection images at finer pixel sampling rates which means that the number of electrons per pixel will be reduced in order to keep the total electron dose from growing. Thus, as the field evolves toward higher resolution, the images and hence the reconstruction will get noisier, making it crucial for methods that process CET data to possess the property of being able to perform well in extremely low SNR environment.



(a)



(b)



(c)

Figure 2.1: Whole cell CET of *S. solfataricus* infected with STIV (a) A 10 nm slice of a 3-D tomogram, computationally sectioned perpendicular to the direction of beam. Multiple pyramid-like protrusions were observed. DNA-free procapsid and DNA-filled virions were clearly identified in the cytoplasm. The quasi-crystalline viral array was labeled. (b) & (c) Enlarged views of some breakage sites in the ruptured cells. The fragments of reminiscent pyramid structures were labeled, which lack S-layer and can be distinguished from typical cell wall. SL, S-layer; CM, cytoplasmic membrane; Pyr, pyramid like protrusion; Pyr Frag, pyramid fragments; IPB, intra-pyramidal body. Scale bar, 200 nm.

	SNR			
	0.1	0.05	0.01	0.005
$\pm 70^\circ$	3.28, 3.53	3.29, 3.58	3.37, 3.65	3.47, 3.94
$\pm 60^\circ$	3.29, 3.55	3.29, 3.57	3.34, 3.73	3.40, 4.21
$\pm 50^\circ$	3.29, 3.56	3.29, 3.61	3.35, 3.74	3.69, 4.22

Table 2.1: Resolution in nm at which that the pair of reconstructed structures are consistent with the pair of reference structures using a cutoff value of 0.5 in FSC. For each experimental condition (SNR and tilt range), a pair of values are reported, which are the resolutions of the reconstructions of the procapsid and capsid structures, respectively.

	SNR			
	0.1	0.05	0.01	0.005
$\pm 70^\circ$	0	1	3	8
$\pm 60^\circ$	0	1	3	10
$\pm 50^\circ$	0	0	6	17

Table 2.2: Number of alignment errors. For each experimental condition (SNR and tilt range), the entry is the number of the 60 data cubes in which the orientation estimate based on the reconstruction is not correct.

	Tilt range		
	$\pm 70^\circ$	$\pm 60^\circ$	$\pm 50^\circ$
alignment error	0.062	0.098	0.206

Table 2.3: Size of alignment errors at SNR = 0.005 as a function of the range of tilt angles for those cubes that are misaligned. The error is quantified as the average of the error present in each pair of true and estimated orientations. The error in a single pair is quantified as the “Euclidean” or “Frobenius” matrix norm [68, p. 358] (the square root of the sum of the squares of the elements in the matrix) of the difference between the rotation matrices which correspond to the true and estimated orientations. Since the particle has icosahedral symmetry, there are 60 equivalent rotational matrices for any orientation and the choice which gives the lowest alignment error is the matrix that is used.

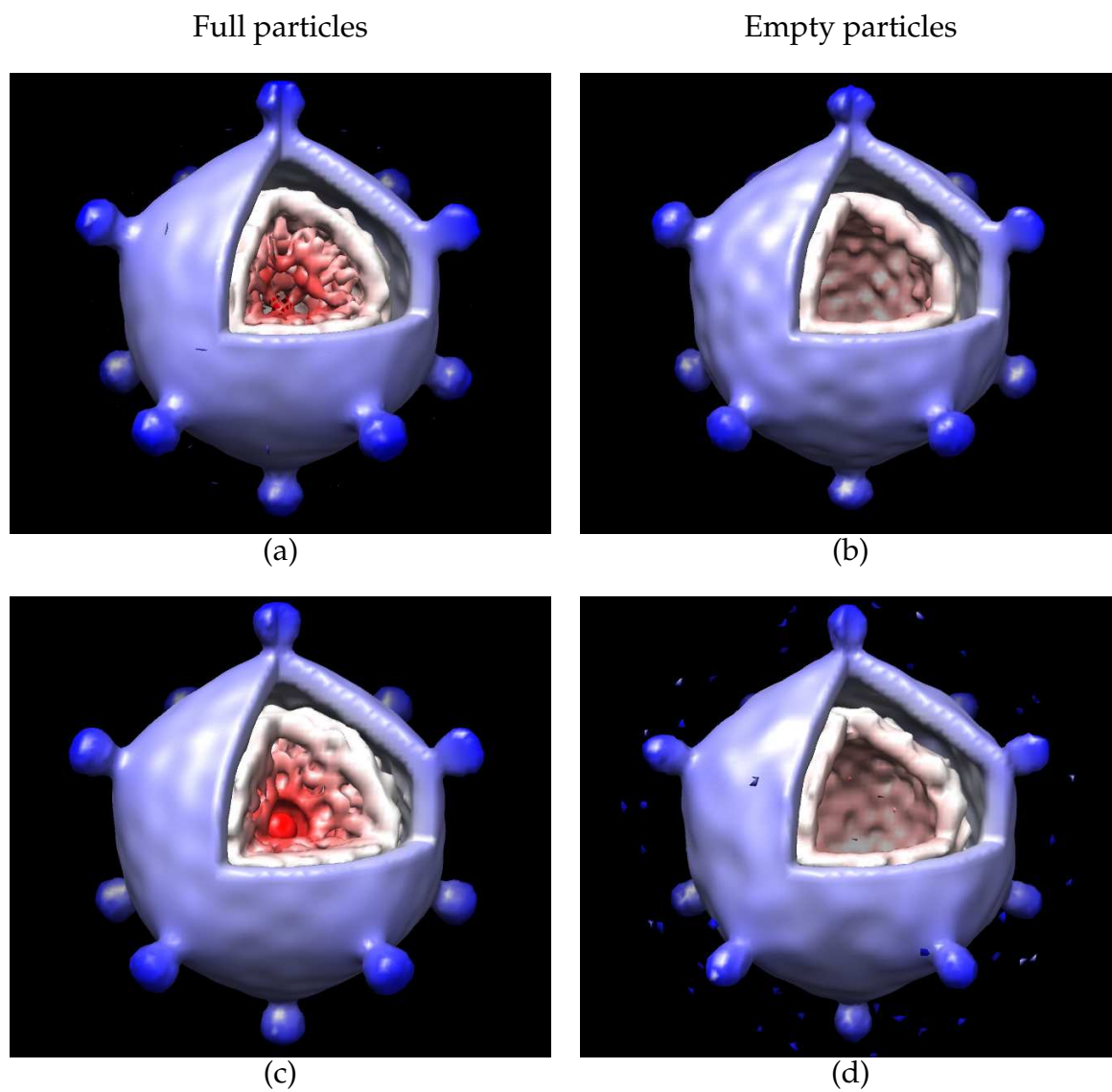
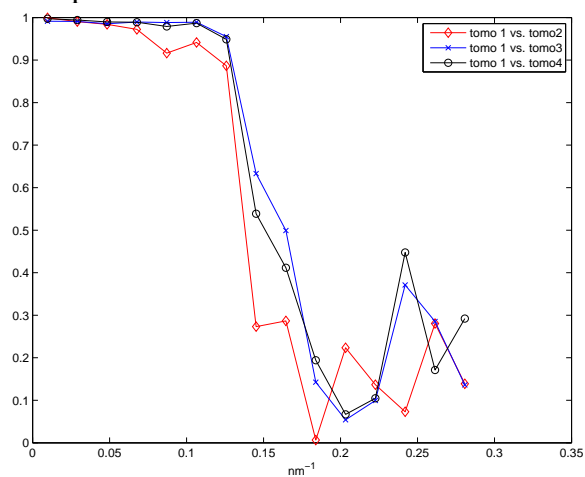


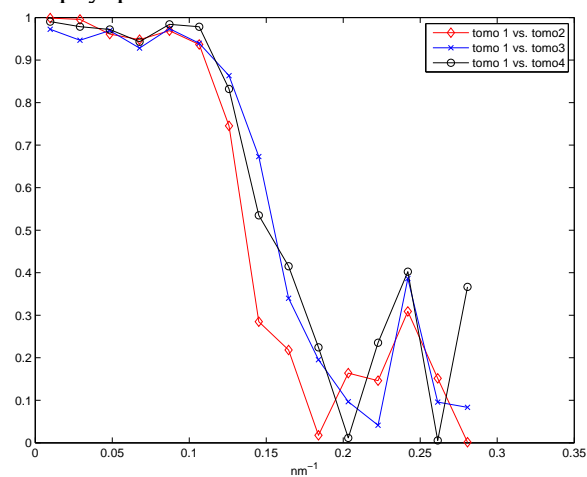
Figure 2.2: Reconstructions of full and empty particles from two of the four tomograms. Panels (a) and (b): First tomogram. Panels (c) and (d): Second tomogram.

all possible FSC curves among the four full-particle reconstructions.



Full particle

all possible FSC curves among the four empty-particle reconstructions.



Empty particle

Figure 2.3: Fourier Shell Correlation (FSC) curves between all pairs of full-particle reconstructions and empty-particle reconstructions.

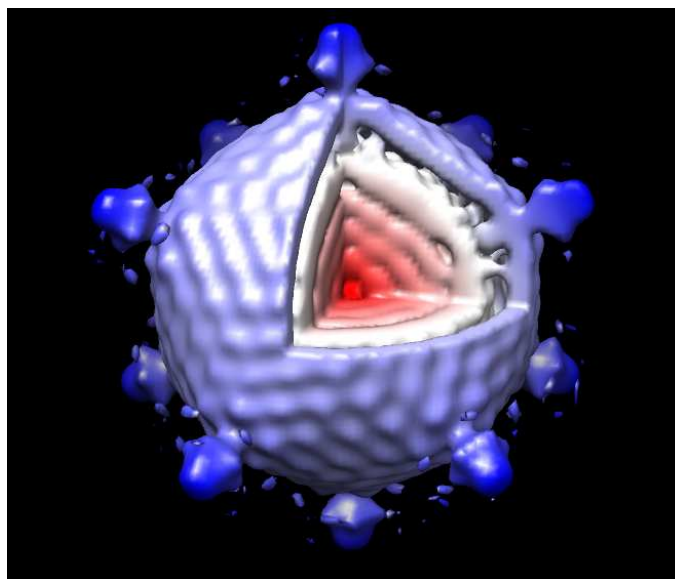


Figure 2.4: Single particle cryo EM reconstruction of STIV from Ref. [58].



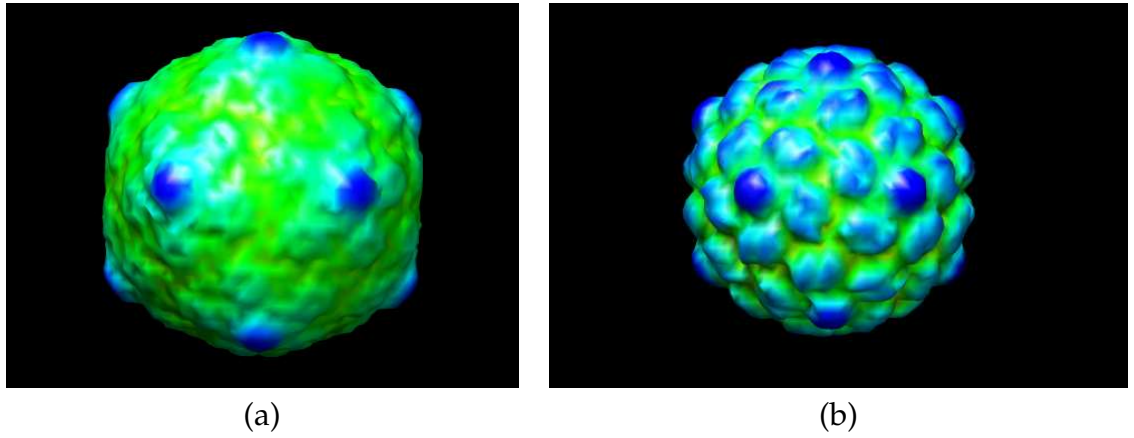


Figure 2.5: 3-D surface rendering by Chimera of the low-resolution reference structures for HK97. Panels (a) and (b) show the mature capsid and the procapsid, respectively.

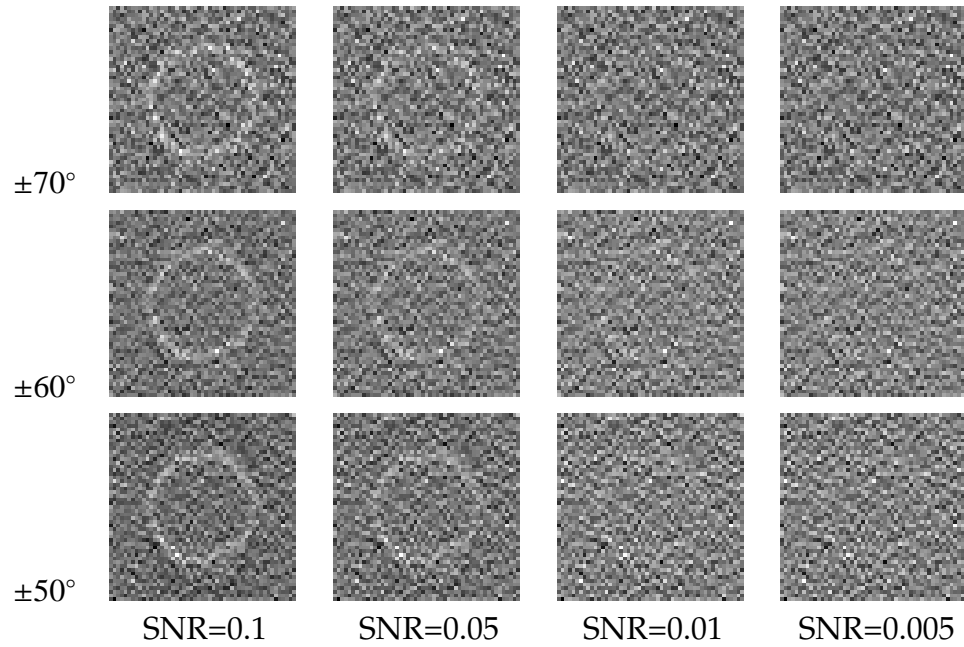


Figure 2.6: Cross-sections of simulated data subcubes of HK97 at various SNR settings and tilt ranges. Each row represent data from same tilt range, which are  $\pm 70^\circ$ ,  $\pm 60^\circ$ , and  $\pm 50^\circ$  from top to bottom. The interval between tilts is always  $2^\circ$ . Each column represent data from same SNR setting, which are 0.1, 0.05, 0.01, and 0.005 from left to right.



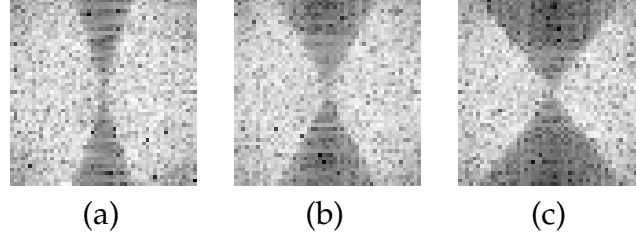


Figure 2.7: Cross-sections in reciprocal space of the simulated data sub-cubes of HK97 for three different tilt ranges. Panels (a), (b), and (c) show tilt ranges of  $\pm 70^\circ$ ,  $\pm 60^\circ$ , and  $\pm 50^\circ$ , respectively.

	Number of data cubes			
	10	20	30	60
resolution	3.45, 4.36	3.45, 4.28	3.45, 4.15	3.34, 3.73
misalignment	2	2	0	3
misclassification	0	0	0	0

Table 2.4: Summary of the second set of experiments with synthetic HK97 CET cubes. Resolutions in nm for the reconstructions of procapsid and capsid are reported in pairs in Row 1 for procapsid and capsid, respectively. Numbers of misaligned and misclassified cubes are reported in Rows 2 and 3, respectively.

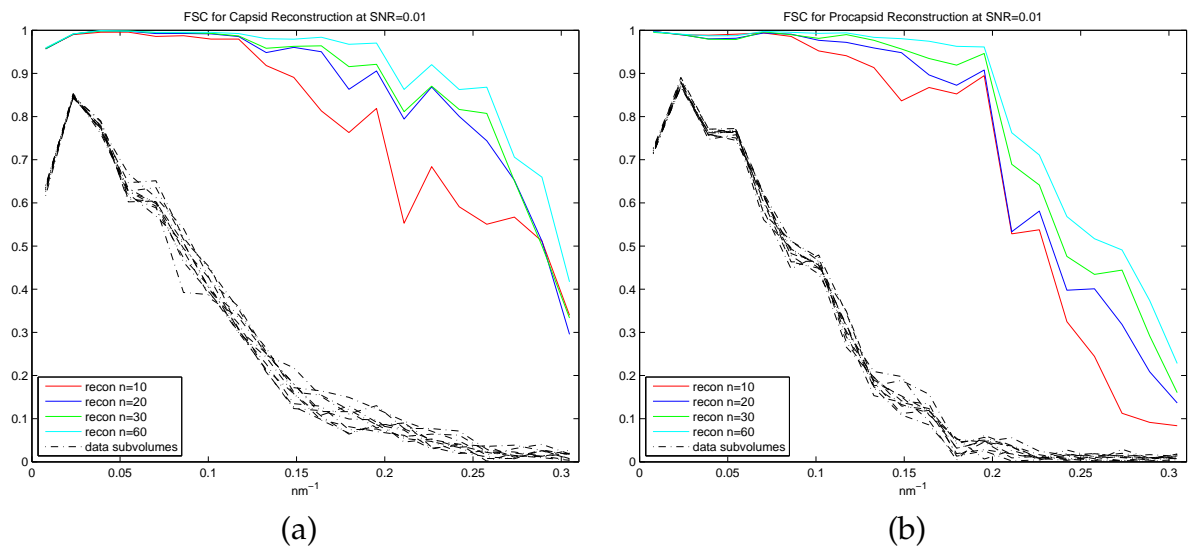


Figure 2.8: Plots of the Fourier Shell Correlation results for correlation between reconstruction and reference structures (colored curves) and for correlation between data cubes and reference structures (black curves). The reconstructions are based on 10, 20, 30, and 60 synthetic HK97 cubes. Each cube is the result of simulating a single-tilt experiment with tilts in the range  $\pm 60^\circ$  and with  $\text{SNR} = 0.01$ . The resolution gain achieved by this algorithm is essentially the difference between the black curves and the colored curves.

## CHAPTER 3

### *IN VIVO* ASSEMBLY OF AN ARCHAEL VIRUS STUDIED WITH WHOLE-CELL CRYO-ELECTRON TOMOGRAPHY

The simultaneously classification and reconstruction algorithm based on maximum likelihood estimator (MLE) is applied to cubes extracted from CET tomograms, each containing a single cell that is infected with STIV virus. The new method is able to identify two distinct structural states of the virus, a procapsid without genome and a full virion, without any initial references from previous studies. By correlating tomograms with these reference structures, we are able to identify structural states, locations, and orientations of all the particles within these tomograms. Quantitative analysis of these particles reveal that virions tended to form tightly packed clusters or quasicrystalline arrays while procapsids mostly scattered outside or on the edges of the clusters. The interesting finding is published in Ref. [35], which we reproduce below. The detail of the methodologies used to quantitatively analyze the viral clusters and arrays are reported in Appendix A.1.

### 3.1 Introduction

To gain mechanistic understanding of virus assembly, it is important to follow those events in the context of the host cell. Cryo-Electron Tomography (CET) is an emerging technique that allows such processes to be imaged in 3D with macromolecular (46 nm) resolution (Grunewald and Cyrklaff, 2006; Koster et al., 1997; Lucic et al., 2005). In brief, cells are preserved in close-to-native states by plunge-freezing (Dubochet et al., 1988; Iancu et al., 2006), a tilt series of projection images are recorded while the sample is kept frozen, and then a 3D re-

construction of the object is calculated by back-projection or other algorithms (Jensen and Briegel, 2007; Lucic et al., 2005).

Here, we applied CET to the intact archaea *Sulfolobus solfataricus* cells infected with *Sulfolobus* turreted icosahedral viruses (STIV). STIV belongs to the PRD1-Adeno lineage of dsDNA viruses as demonstrated by the structural similarity of the major capsid protein and the overall capsid architecture in this lineage (Khayat et al., 2005). The icosahedral capsid is 74 nm in diameter, has a pseudo  $T = 31d$  quasiequivalent surface lattice and an inner membrane that encloses a 17.3 kbp circular dsDNA genome. At the twelve five-fold vertices are turret-like appendages extending 13 nm above the capsid shell (Rice et al., 2004). Mass spectrometry studies of purified virions suggest that the inner membrane contains several viral gene products and shows the presence of acidic tetraether lipids corresponding to an enrichment of a subpopulation of lipid found in the host (Maaty et al., 2006).

The STIV replication cycle can be followed using a near synchronous, single cycle infection of *S. solfataricus* (strain P2) which grows optimally at 80°C and pH 3 (Ortmann et al., 2006). *S. solfataricus* has an oval shape about 1 µm in diameter. Exterior to the cytoplasmic membrane is a surface protein layer (S-layer) that is formed by a single glycosylated protein (Engelhardt and Peters, 1998; Grimm et al., 1998). Microarray analysis and transmission EM (TEM) of thin sections demonstrated that the transcription of the STIV genome peaks at around 24 hr postinfection (hpi) and cells containing assembled viral particles become dominant at around 32 hpi (Brumfield et al., 2009; Ortmann et al., 2006). TEM and scanning EM (SEM) analysis of cells infected by STIV show pyramid-like structural protrusions from the cell surfaces (Brumfield et al., 2009). These

protrusions appear to be the sites where infected cells lyse at the late infection stage in order to release virions (Brumfield et al., 2009). Some viral particles shown in TEM did not appear to contain genomes and were proposed to be procapsids. Although possible artifacts, caused by negative stain sample preparation, could not be ruled out, the results suggested the existence of a procapsid prior to the formation of a genome containing virion (Brumfield et al., 2009).

We plunge-froze *S. solfataricus* at different time points postinfection. The whole-cell tomograms showed STIV particles in different states of assembly. A maximum likelihood (ML) algorithm was developed to automatically identify, classify and reconstruct two classes of viral particles within the cell. The reconstructions were used as models to identify and orient particles in other tomograms based on a template matching approach (Frangakis et al., 2002). This procedure made it dehydration and staining procedures. Pyramids started to form before viral particles became visible in the cytoplasm as seen in the samples at 22 and 24 hpi. Figure 1D shows some pyramids in the process of protruding out of a thinning cell wall and perturbing the S-layer. Electron dense bodies with diameters of 90-110 nm were often found inside the pyramids and are referred to as intrapyramidal bodies (IPBs) (Fig. 3.1A and B). Electron dense bodies were also observed in random locations in uninfected cells; however, they were almost always associated with pyramids in infected cells. Some IPBs appeared to have extra layers of material wrapped around them that were less dense than the cores (Fig. 3.1B, bottom panel). The compositions and functions of IBPs and whether they are different from those observed in the uninfected cells remain to be characterized.

## 3.2 Assembly and Maturation of STIV Observed In Vivo

Particles with icosahedral shape and dimension similar to STIV could be clearly identified in the cytoplasm (Fig. 3.1A). Full and empty particles were readily distinguished (Fig. 3.2 A and B). The radial density plots illustrated the differences in internal density in full and empty particles (approximately 0-15 nm radius from the centers), which corresponded to the packed DNA genome as characterized by single particle reconstruction of STIV virions (Fig. 3.3 (d) ) (Rice et al., 2004). Both particles types had an outer capsid shell ( $\sim 37.5$  nm in outer radius) and an inner lipid membrane ( $\sim 25$  nm in outer radius), comparable to what was seen in the single particle reconstruction (Fig. 3.3 (d)). These observations confirm that the STIV life cycle involves a procapsid (empty particles) where the capsid and membrane are assembled prior to genome packaging.

Detailed 3D reconstructions of the STIV virions were needed to show the changes that accompany virus maturation in vivo. Electron cryotomograms are challenging to interpret due to the missing wedge artifact and low signal-to-noise ratio. To fill in the missing data and boost the signal-to-noise ratio, particles from multiple subtomograms were aligned and averaged. An average 3D reconstruction of the particles was produced with a multiclass ML algorithm from 123 subvolumes containing one STIV particle each. The algorithm, as it was applied, allows for multiple reconstructions to be derived simultaneously (Fig 3.3 (a) and (b) ). Following a detailed analysis of these reconstructions, there were only two distinct structures that emerged: empty and full. If other subtle differences in particle structure exist, the data were not of sufficient resolution to resolve them. More detailed information can be found in the methods. The three perpendicular cross-sections of the averaged volumes display turrets at

5-fold vertices, indicating the correct computation of orientations for each sub-volume. The resolution was estimated to be approximately 6.5 nm based on the Fourier Shell Correlation (FSC) method with a 0.5 threshold. The radial density plots for the procapsid and virion reconstructions displayed overall agreement for the capsid and membrane radii with each other and with the single particle reconstruction (Fig 3.3 (d)). Comparison of the full and empty particles demonstrate that no large-scale reorganization of the capsid or membrane occur upon DNA packaging in STIV. The appearance of turretlike structures at 5-fold vertices in the reconstructions did not result from model-bias since the ML approach is *ab initio*. Both virions and procapsids had closely similar turret-like structures, showing that they were assembled on the particle prior to genome packaging. A three-class reconstruction was also carried out (data not shown). The resulting reconstructions display one DNA-free and two DNA-containing reconstructions. However, no significant differences in the two DNA-containing reconstructions can be concluded. A three-class reconstruction cannot identify partially packaged particles if there are any.

In addition to virions and procapsids, partially assembled shells formed by capsid and membranes were also observed (Fig. 3.2 C). These partial assemblies have curvature and spacing between capsid and membrane layers closely similar to those of fully assembled procapsids. They were not associated with the cytoplasmic membrane and clearly did not form by a budding mechanism. The data indicate that the STIV capsid and membrane coassemble in the cytoplasm to form a procapsid, which further matures to a virion through a DNA packaging event.

### 3.3 Distribution and Packing of Viral Particles In Vivo

In order to robustly classify viral particles and determine their locations and orientations in a large number of tomograms, without applying the computationally intensive ML process for each tomogram, a correlation-based approach was employed using the reconstructions derived by the ML method as references. More detailed information can be found in the methods. A gallery of cells sampled at 29-32 hpi is shown in Figure 4A. Each cell has between 22 and 166 particles. Viral clusters or arrays were frequently present and contained the majority of particles in the cell (Table 3.1). Cells that did not have viral clusters had low particle number (22 - 44) and a high ratio of empty procapsids to virions (55% - 73%), indicating that these cells were at an early stage of infection. Particle clusters mainly consisted of virions. Procapsids were only found on the edge of the clusters or not associated with the clusters at all. Particles in clusters were packed very tightly as illustrated in the blow-up view of a cluster (Fig. 3.5). The narrow distribution of distances between the centers of particles and their closest neighbors are shown in Fig. 3.6 (a). The average center-to-center distance between adjacent particles is 75 nm, corresponding to the particle diameter without the turrets. This is consistent with the orientation analysis that demonstrates the turrets of nearest neighbors avoid each other in the arrays (Fig 3.5; Fig 3.6 (b)). Some viral clusters had close-packed lattices with quasicrystalline, interparticle spacing as illustrated in Fig. 3.7 (a)-(d).



### 3.4 Discussion

STIV infection induces dramatic host morphological changes. The pyramid structures are unlike cellular protrusions typically observed in eukaryotes (Charras and Paluch, 2008; Mattila and Lappalainen, 2008). Based on what was characterized by CET, we propose a model for pyramid formation (Fig 3.1E). The pyramids could form by either mechanically protruding and/or enzymatically digesting through the S-layer, perturbing the S-layer integrity and detaching it from the pyramid membrane. The thicker pyramid membrane and sharply defined facets indicate different protein and/or lipid compositions from the cytoplasmic membrane, implying the transport of specific proteins and extra lipids to the target sites to build pyramids. The compositions and functions of IPB remain to be characterized. The fact that not every pyramid has IPB and pyramids at an early stage of formation do not have IPB suggests that they may not relate to the pyramid formation or growth but likely relate to some pyramid-mediated viral release mechanism.

The control mechanisms that assemble a viral capsid with hundreds of copies of repeated subunits are yet to be fully understood (Caspar, 1980; Caspar and Klug, 1962; Fane and Prevelige, 2003; Johnson and Speir, 1997). The issue is further complicated in viruses that contain inner membranes. Outstanding questions regarding these processes include: (1) How does the membrane core assemble? Does it form independently, followed by capsid assembly or do capsid proteins and membranes coassemble? (2) Where does viral lipid originate in cells that lack membrane-containing organelles; i.e., does the membrane core bud off the cytoplasmic membrane or form *de novo* in the cytoplasm? This CET study of STIV provided valuable insights of *in vivo* assembly and explic-

itly addresses some of these points. The observation of partially built particles supports the hypothesis that assembly of the STIV capsid shell and the membrane are tightly coupled. The curvature and layer spacing of partial shells resembled those of fully assembled procapsids, implying defined local interactions between capsid and membrane and between capsid subunits as assembly proceeds. There is no evidence from our CET study that partially assembled particles bud off the cytoplasmic membrane. Indeed, the mass spectrometry study concluded that the lipid compositions of viral and cellular membranes are different (Maaty et al., 2006).

The crystal structures of PRD1 and PM2, viruses in the same lineage as STIV, revealed how proteins and membranes organize in virions (Abrescia et al., 2004, 2008; Cockburn et al., 2004). About half of the membranes mass is attributable to proteins, which is likely to be the case for the STIV internal membrane. The crystal structures reported suggest that the membrane-associated proteins may function as tape-measures or scaffolding proteins, similar to those in viruses that lack inner membranes (Abrescia et al., 2004, 2008; Cockburn et al., 2004; Fane and Prevelige, 2003). STIV gene products may embed in the nascent lipid creating curvature and affinity for the major capsid protein. In a previous study, the C-terminal region of the STIV capsid protein was shown to interact with membrane (Khayat et al., 2005). We propose that the viral lipids derive *de novo* and associate through their hydrophobic nature. The association of trans-membrane proteins organizes the membrane through specific intersubunit interactions that facilitates correct binding interactions of capsid proteins as assembly continues (Fig 3.8).

The CET study showed that the STIV life cycle involved a procapsid, as ob-

served in PRD1 (Martin et al., 2001), human adenovirus (Ostapchuk and Hearing, 2005), as well as dsDNA bacteriophages (Casjens and King, 1975). By comparing the reconstructions of the procapsid and the virion, it is clear that DNA packaging does not induce capsid conformational change that can be resolved at 6.5 nm resolution. At the current resolution, subtle changes cannot be detected. For instance in PRD1, single particle reconstructions show a subtle reduction of separation between the capsid and membrane due to a 5% radial expansion and 10% thinning of the membrane when the genome was encapsidated (Butcher et al., 1995; Martin et al., 2001).

The 5-fold vertices of icosahedral viruses are often found to be involved in receptor binding and/or genome translocation (Abrescia et al., 2004, 2008; Gowen et al., 2003; Johnson and Chiu, 2007; Moore and Prevelige, 2002). The presence of turret-like densities at the 5-fold vertices in both virions and procapsids indicates that the vertex proteins associate while procapsids assemble. It remains to be determined whether any of the 12 turrets can function in DNA packaging or if there is a specialized vertex for DNA packaging like in PRD1, Adenoviruses, Herpes virus, and dsDNA phages (Bazinet and King, 1985; Cardone et al., 2007; Chang et al., 2007; Karhu et al., 2007; Ostapchuk and Hearing, 2005).

We propose that crystalline viral arrays offer an important advantage as they allow the limited cellular space to accommodate the greatest number of viruses. In fact, virus arrays have been observed in other systems such as in papillomaviruses (Campo, 2002; Wang et al., 2009), FHV (Lanman et al., 2008), and iridovirus (Darlington et al., 1966). The high image quality of cellular CET and the robust computational analysis allowed the STIV arrays to be analyzed in

exceptional detail. Virus factories or viroplasm have been reported in eukaryotic and prokaryotic systems where proteins and newly synthesized genomes are confined within specific compartments for efficient viral replication and assembly (Bravo et al., 2005; Cook, 1999; Netherton et al., 2007). The analysis of the distribution of STIV virions and procapsids in multiple cells suggests that genome packaging leads to redistribution of particles. The viral clusters of STIV may accommodate DNA and packaging enzymes where capsid assembly and genome packaging are tightly coupled, accounting for the observation that procapsids appear only on the edge of the clusters. Alternatively, particles may assemble outside the clusters and traffic to specific sites where viral genome packaging takes place with the formation of tightly packed clusters.

STIV infected *sulfolobus* provided an ideal system for the CET study of complex virus assembly and maturation as well as raising intriguing questions about virus-induced cellular changes. Future studies to elucidate greater detail regarding the proteins directing and participating in the all aspects of the virus life cycle will utilize specific gold-labeled antibodies to cellular and viral gene products as well as an available infectious clone for STIV that allows the mutation or deletion of specific viral genes.

## **3.5 Experimental Procedures**

### **3.5.1 Sample Preparation**

*S. solfataricus* strain 2-2-12 cultures were infected with STIV as previously described (Ortmann et al., 2006). *Sulfolobus* were grown up from glycerol stock in

media 182 (pH 3.5), 80° C and passage once. The cultures were passed second time to media 182 (pH 2.5) and infected with STIV at MOI ~ 2 at OD 650 nm ~ 0.4. At 22–32 hr postinfection, cells were concentrated 20 fold by centrifugation. Equal volume of cells were mixed with BSA-treated 10 nm gold colloidal beads and applied to plasma cleaned Quantifoil holey carbon films. The grids were plunge-frozen in liquid ethane using a Vitrobot (FEI Inc.) and stored in liquid nitrogen (Iancu et al., 2006).

### **3.5.2 Tomography Data Collection and Image Processing**

Specimens were imaged using a FEI Polara transmission electron microscope at an accelerating voltage of 300 kV. Tilt series covering an angular range from –60 to +60 with 1° increment were acquired automatically using Legion at 18,000x magnification, 10-14  $\mu\text{m}$  underfocus, with 110-160 electrons /  $\text{\AA}^2$  total dose (Suloway et al., 2005, 2009). Images were recorded with a lens-coupled Gatan UltraCam 4k 3 4k CCD camera binned by two so the final pixel size represented 1.26 nm on the specimen. Tilt series were aligned with gold fiducials and 3D reconstructions were calculated with IMOD (Kremer et al., 1996). Tomograms were filtered with a 3D median filter with 3dmod and visualized and/or segmented using 3dmod and Chimera (Kremer et al., 1996; Pettersen et al., 2004). Totally about 60 tomograms were analyzed.

### 3.5.3 Maximum Likelihood Reconstruction

The icosahedrally-symmetric reconstructions were computed from the tomography volumes by a modification of the single particle cryo-EM reconstruction algorithm (Prust et al., 2009; Yin et al., 2003). First 123 subvolumes, each containing one STIV particle, were extracted from the tomography by crosscorrelating with a spherically symmetric template. Then the modification of the algorithm of (Yin et al., 2003) was applied to jointly compute two 3D icosahedrally symmetric reconstructions, which correspond to full and empty particles. Third, the STIV particle shown in each of the subvolumes used in the reconstruction was classified with respect to being a full or an empty particle. The modification of the ML reconstruction algorithm (Yin et al., 2003) takes into account the following aspects of electron tomography. In Yin et al. (2003), the cryo-EM image is a linear transformation of the 3D distribution of electron scattering intensity and the transformation includes projection from 3D to 2D with unknown projection angles, translation of the location of particle center (two component vector), and the contrast transfer function. Now the transformation includes rotation of the particle in the subvolume, translation of the location of the center of the particle in the subvolume (three component vector), and the fact that a wedge of data in 3D reciprocal space is missing due to the fact that the tilt angle is limited to  $\pm 60^\circ$ . To estimate the resolution, the whole data set of 123 subvolumes was separated into two subsets. For each data set, reconstructions assuming two classes of particles are performed. Then, for each class, Fourier Shell Correlations are calculated between reconstructions from the two different subsets. Using a cutoff of 0.5, the resolutions for Class 1 and class 2 reconstructions are 6.5 nm.

### 3.5.4 Template Matching Approach

The reconstructions were used as references to determine particle classes (i.e., full versus empty), locations, and orientations in the second cell tomogram. A correlation-based approach was used as described in the following steps. First, a one-class icosahedrally-symmetric reconstruction was computed as described previously in this chapter and spherically averaged to provide a template. Second, the template was used to locate particle centers in the tomograms by cross-correlation. Third, subvolumes of  $65 \times 65 \times 65$  voxels were extracted from the cell tomogram based on the center locations. Fourth, the icosahedrally-symmetric reconstructions obtained in the two-class reconstruction were used to generate templates to determine particle orientations. For each reconstruction, a library of 5000 templates was constructed to cover the reconstruction in different orientations described by Euler angles ( $a, b, g$ ) (all within three adjacent fundamental domains of the icosahedral group). The missing wedge was taken into account when generating templates. Fifth, using the 10,000 templates from step four, the subtomogram centered at each location determined in step two and extracted in step three is both labeled and orientated by normalized correlation. Then, based on these estimated orientations and class labels, each particle is aligned and averaged within its class with the application of the icosahedral symmetry operations. The maximum peak value of the correlation attainable over orientations at each SNR was simulated by correlating the templates and the templates plus white Gaussian noise.

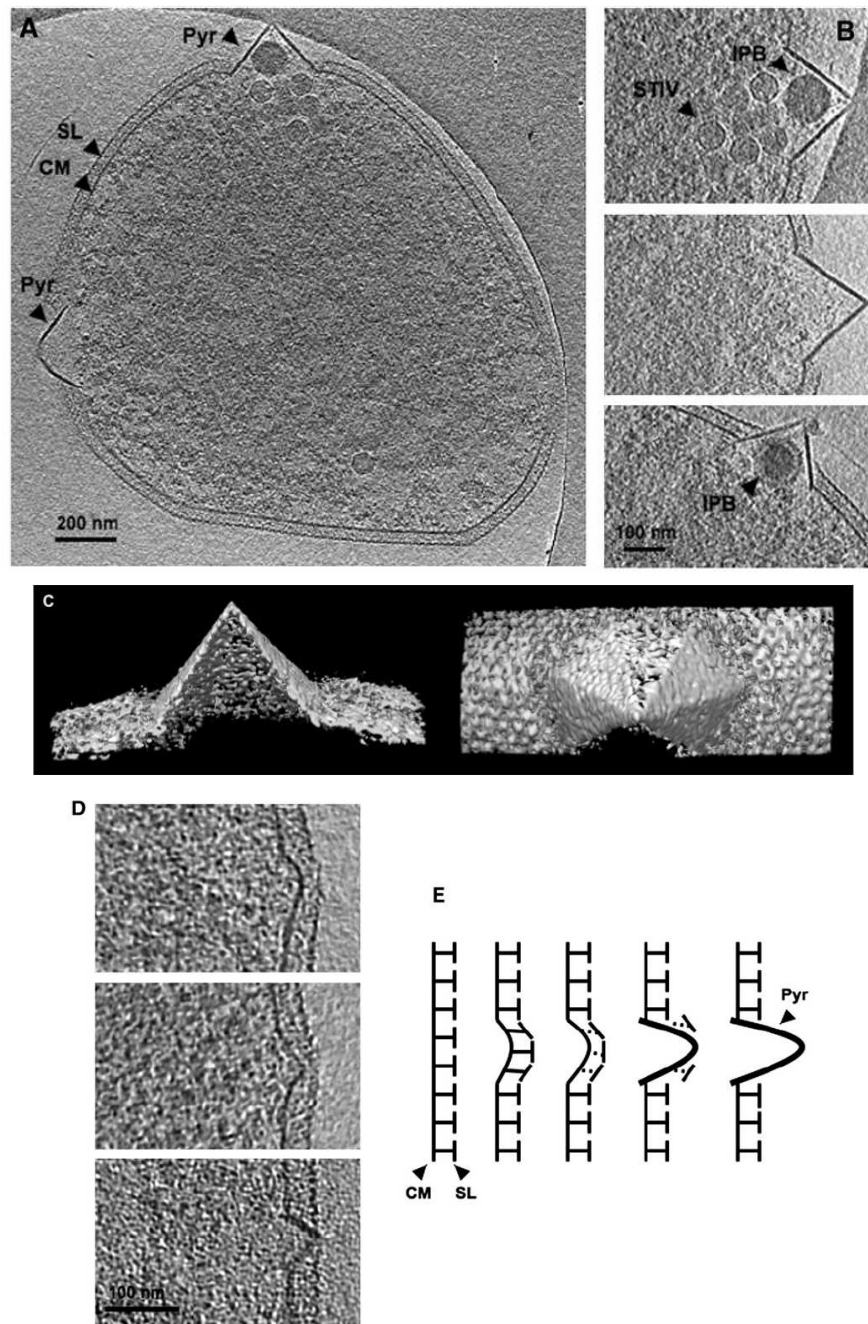


Figure 3.1: Whole-cell CET of *S. solfataricus* infected with *STIV*: (A) A 20 nm slice of a 3D tomogram, computationally sectioned perpendicular to the direction of the beam (B) Enlarged views of some representative pyramids. (C) Surface representations of a pyramid viewed from the side and the top of the structure. (D) Enlarged views of some pyramids at early stages of formation. (E) A model of pyramid formation. A pyramid forms by either mechanically protruding and/or enzymatically digesting through the S-layer. The S-layer structure is perturbed and finally detaches from pyramid membrane. Specific proteins and lipids are recruited as a pyramid builds. SL, S-layer; CM, cytoplasmic membrane; Pyr, pyramid like protrusion; STIV, STIV particles; IPB, intrapyramidal body. Scale bar, 200 nm (A), 100 nm (B and D).



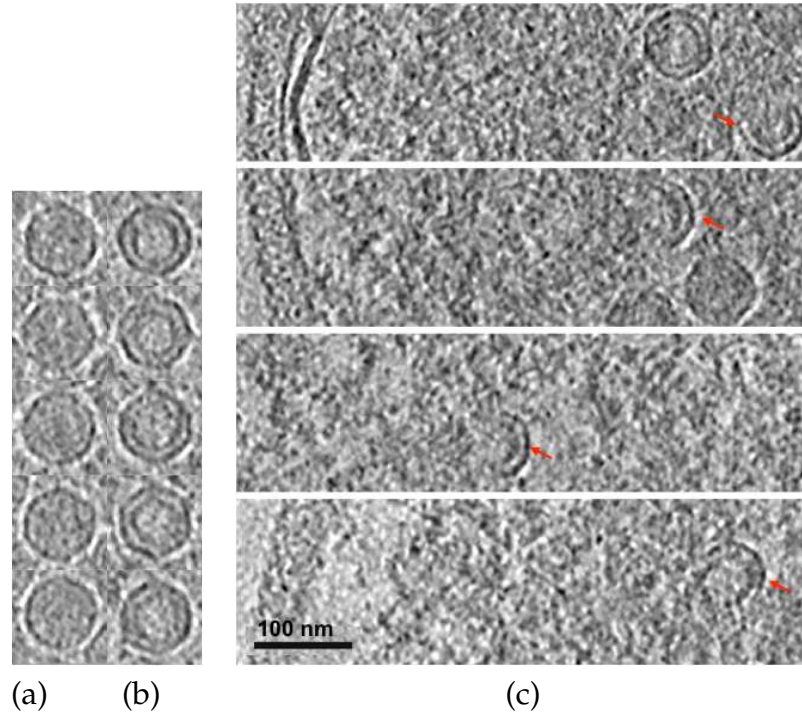


Figure 3.2: STIV Virions, Procapsids, and Partially Assembled Particles Observed In Vivo Subtomographic slices displaying STIV virions (A) and procapsids (B) and partially assembled particles (C) that contain parts of capsid and membrane viewed perpendicular to the direction of the beam (x-y plane). Scale bar, 100 nm (C).

Total number of particles:	22-166
Percentage of virions:	27% - 86 %
Percentage of particles in viral clusters:	67% - 99 %
Percentage of particles in viral clusters being virions:	74% - 94%

Table 3.1: The analysis of viral distribution and array packing

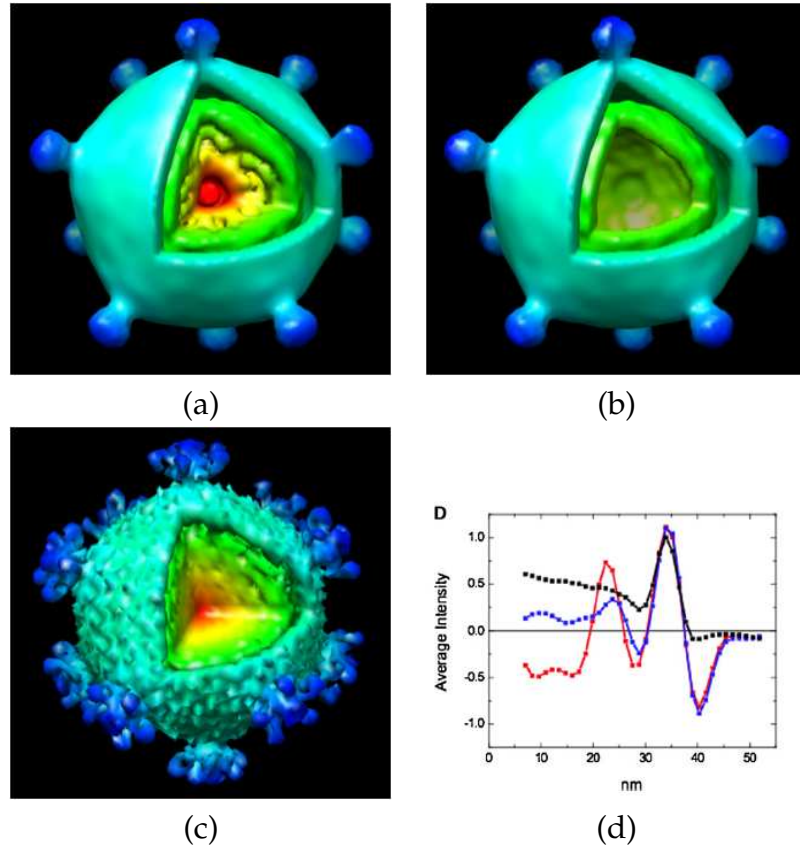


Figure 3.3: The Reconstructed Density of the Two Classes of STIV Particles Determined by the ML Algorithm. The reconstructions of a virion (A) and a procapsid (B) determined with 123 particle-containing subvolumes of the cellular tomogram. (C) The surface representation of the single molecule reconstruction of purified virions. (D) The radial density plots of the subtomographic reconstructions of the virion (blue), procapsid (red), and the single molecule reconstruction of purified virions (black).

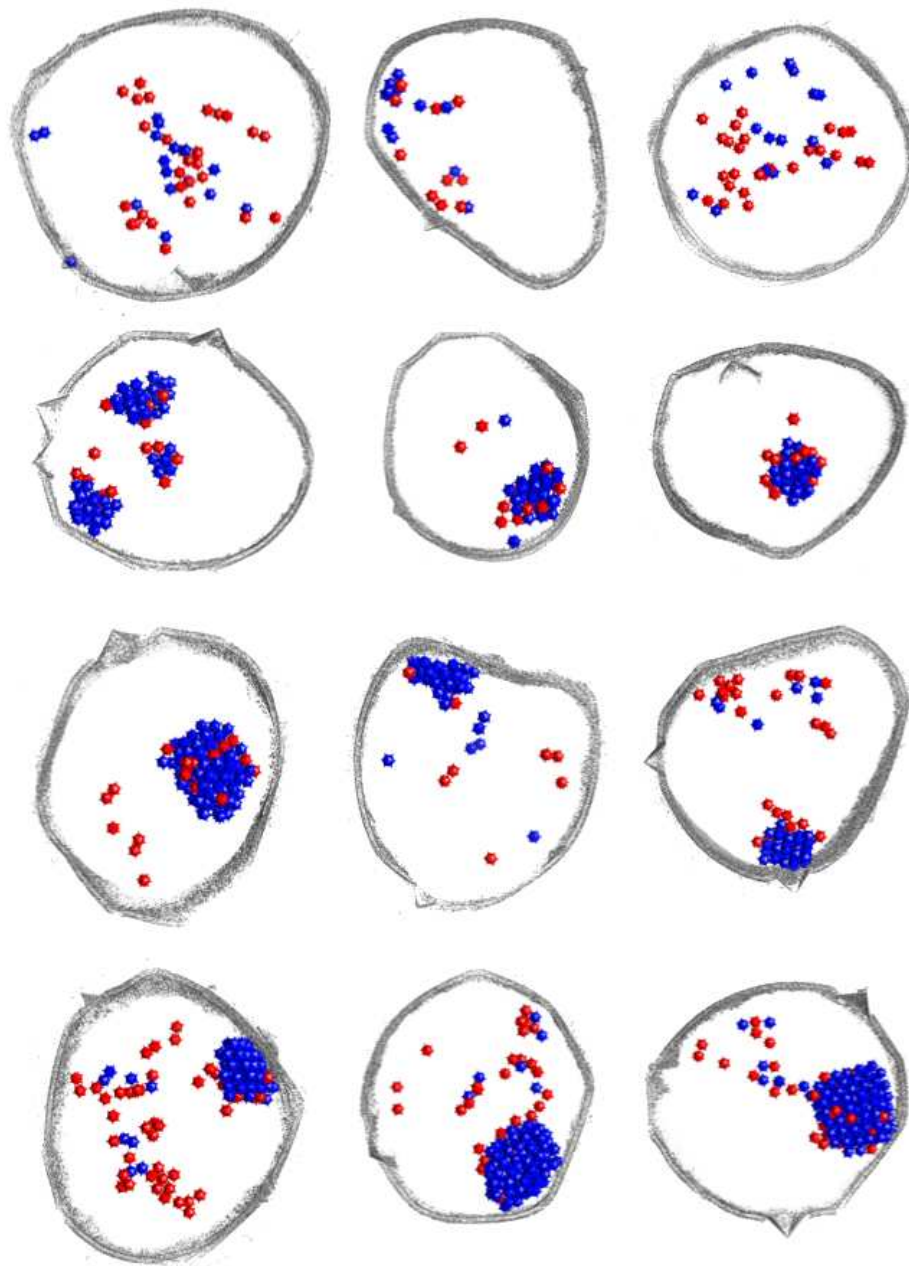


Figure 3.4: The Analysis of Viral Distribution and Packing: A gallery of model representations of viral distributions with cell periphery outlined in gray.

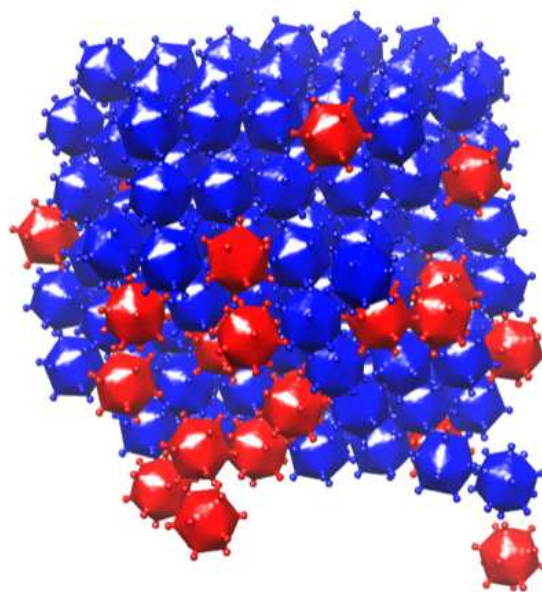


Figure 3.5: The Analysis of Viral Distribution and Packing: The model representation of a quasicrystalline packing of a viral array (virions, blue icosahedra; procapsids, red icosahedra)

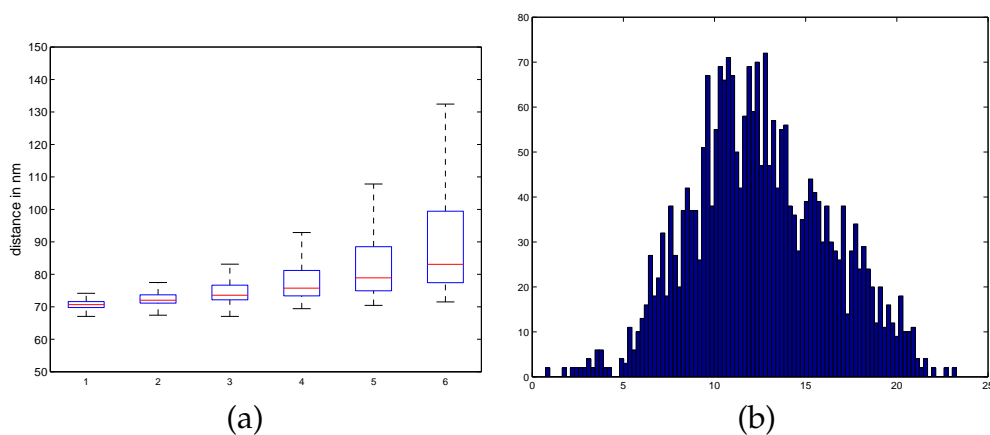


Figure 3.6: (a) The box plot of the distance between the centers of the particles in clusters to its 1st to 6th closest neighboring particles. Red bar, median; Blue box, range covered by 25-75% of particles; Error bar, minimum and maximum in the distribution. (b) The histogram of the distance between the turrets of its neighboring particles within 75 nm

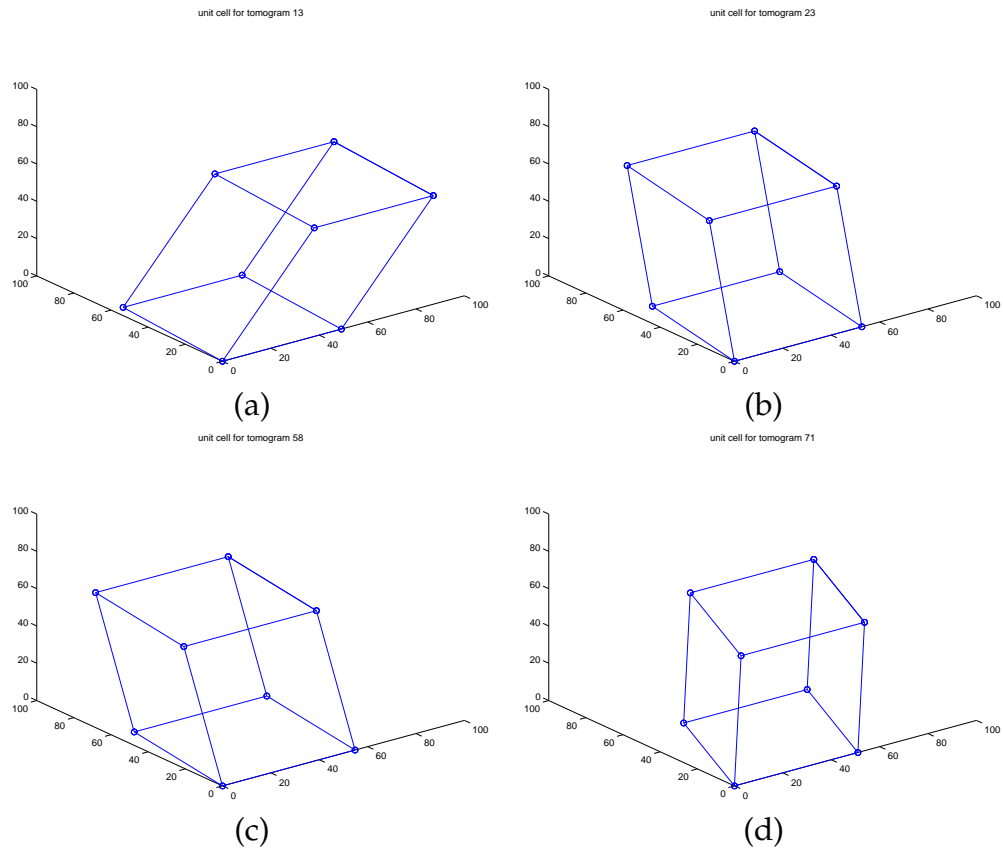


Figure 3.7: The diagrams of unit cells that describe the packing of observed viral arrays

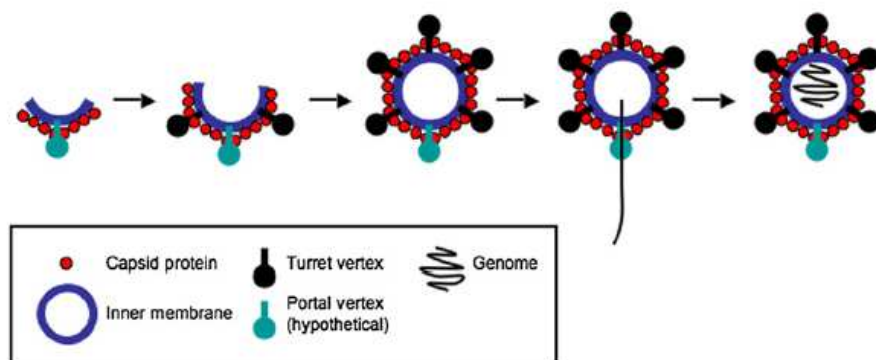


Figure 3.8: A Model of STIV Assembly and Maturation The lipids and trans-membrane proteins assemble to form viral membrane as the trans-membrane proteins serve as tape-measure scaffolding and facilitate correct assembly of capsid proteins. The growing membrane and capsid form a procapsid. The genome is packaged through either a turret vertex or a specialized portal vertex. A procapsid matures to a virion without undergoing large-scale capsid transformation.

## CHAPTER 4

### ASYMMETRY DETECTION WITHIN SYMMETRIC STRUCTURES

#### 4.1 Introduction

Many spherical virus particles are described as having icosahedral symmetry. However, many have an intrinsic asymmetry in the capsid, such as the portal through which dsDNA is loaded into a phage before the attachment of the tail. Others, such as *Sulfolobus* Turreted Icosahedral Virus (STIV) [93, 58, 15, 35], are suspected to have similar asymmetry. Finally, even a particle that has no capsid asymmetry, e.g., Flock House Virus (FHV) [71], probably has an asymmetry in the capsid-genome system because, for example, a mono-partite genome has two distinct ends rather than the 60 identical copies of each distinct end required for icosahedral symmetry so, especially for infection, the system is likely to behave in an asymmetrical fashion. In many systems, e.g., phage, STIV, and FHV, the asymmetry is expected to be located at a 5-fold vertex.

Unless the asymmetry is large in terms of electron scattering intensity, as in recent infectious phage reconstructions from cryo-EM images [69, 54], it has been difficult to detect this asymmetry in x-ray crystallography or cryo-EM data. We propose to study asymmetry of icosahedrally-symmetric particles by Cryo-Electron tomography (CET) of purified particles in the case where the location of the asymmetry can be hypothesized, e.g., a 5-fold vertex. The advantage of CET cubes over cryo-EM images is that regions of interest (ROIs) are not superimposed by a projection operation. We search for asymmetry within a particle by processing these ROIs from each CET cube with a classification and reconstruction algorithm based on maximum likelihood ideas.

Maximum likelihood algorithms depend on a probabilistic description that relates the data to the unknown parameters. A description we have used successfully in the past is that the data is the sum of a linear transformation of the unknown parameters plus Gaussian noise. We modify this standard description with a novel description based on adapting normalized correlation, which is widely used to compare two CET cubes, to maximum likelihood.

Suppose that there is at least a subset of the particles for which the asymmetry is replacement of one region of interest (ROI) from a set of symmetry-related ROIs by a different structure, e.g., the portal of a dsDNA bacteriophage occupies one of the twelve otherwise symmetry-related 5-fold axis locations. In this case, using results from the maximum likelihood estimator, we present methods to compute an asymmetric reconstruction from the CET cubes and methods to use the asymmetric reconstruction to absolutely orient cryo-EM images from which a higher-resolution asymmetric reconstruction can then be computed by established methods.

## 4.2 Maximum Likelihood Estimator (MLE) based on Normalized Correlation

We introduced [24] a statistical model for cryo-EM images based on the equation  $y_i = L(z_i)c_i + w_i$  where  $y_i$  is the  $i$ th image represented as a vector,  $L(z_i)$  describes the electron microscope imaging system where  $z_i$  are the values of the so-called nuisance parameters for the  $i$ th image (e.g., the unknown orientation of the  $i$ th particle in the microscope),  $c_i$  is the unknown vector of parameters for the  $i$ th particle which are the same for all particles in a particular class, and



$w_i$  is the noise (zero mean, covariance  $Q$ ) in the  $i$ th image.  $L(z_i)$  contains the projection operation  $[\mathcal{P}(z_i)]$ , the translation operator which describes the location of the projected center of the particle in the image  $[\mathcal{T}(z_i)]$ , and the effect of the contrast transfer function ( $G$ ), specifically,  $L(z_i) = G\mathcal{T}(z_i)\mathcal{P}(z_i)$ . The model was used as the basis of a maximum likelihood approach, implemented by an expectation-maximization algorithm, to jointly determine a reconstruction for each class from a set of images where the class of particle shown in a particular image is unknown [24, 119, 118, 72, 91]. This approach is now a part of a standard cryo-EM system [98]. We modified the approach to use CET cubes as data [35, 113]. In the modified approach,  $L(z_i)$  contains the orientation of the particle (6 degrees of freedom) in the CET cube  $[O(z_i)]$  and the effect of the missing wedge ( $W$ ), specifically,  $L(z_i) = WO(z_i)$ , and  $w_i$  is voxel rather than pixel noise. A key feature of the modified  $L(z_i)$  is that it includes the fact that a wedge of data is missing in reciprocal space because of the inability to record data at tilts beyond approximately  $\pm 60^\circ$ . Using the maximum likelihood approach, we have already demonstrated the ability to jointly classify, orient (up to an operation from the icosahedral group), and reconstruct (symmetric reconstructions) virus particles from CET cubes of STIV infected *Sulfolobus sulfataricus* cells [35, 113]. under extreme low SNR ( $< 0.01$ ) environment. We propose to develop this approach further and compare it with a novel maximum likelihood estimator based on normalized correlation which is described in the following paragraph.

In comparing two reciprocal-space CET volumes  $a$  and  $b$ , the normalized correlation

$\bar{r} = \sum_{i \notin W} a_i b_i / \sqrt{[\sum_{i \notin W} a_i^2][\sum_{i \notin W} b_i^2]}$ , where  $W = W_a \cup W_b$  and  $W_a$  ( $W_b$ ) is the missing wedge for  $a$  ( $b$ ), has been useful [31]. In a maximum likelihood estimator, the need is to compare one CET volume with a prediction of the contents of the

volume where the prediction depends on the values of parameters. We propose to define

$$m(y_i|z_i, c) = \left\| \frac{y_i}{\|y_i\|} - \frac{L(z_i)c}{\|L(z_i)c\|} \right\|^2 = 2(1 - \bar{r}^2) \quad (4.1)$$

where the norm  $\|\cdot\|$  includes only those elements of  $L(z_i)c$  or equivalently  $y_i$  which are outside of the missing wedge and  $\bar{r}$  is the normalized correlation between the data  $y_i$  and the prediction  $L(z_i)c$ . In both the cryo-EM and CET approaches defined in the previous paragraph, note that the probability density function (pdf) on  $y_i$  conditional on  $z_i$  and  $c$  is Gaussian and proportional to the exponential of a quadratic form, specifically,  $\exp\left(-(1/2)\|Q^{-1/2}(y_i - L(z_i)c)\|^2\right)$ . We propose to investigate the maximum likelihood estimator that results from changing the conditional pdf from  $\exp\left(-(1/2)\|Q^{-1/2}(y_i - L(z_i)c)\|^2\right)$  to  $\exp(-m(y_i|z_i, c))$ . An important characteristic of this approach is that we are able to do pattern classification using features that are not rotationally invariant even though we do not know the correct orientation. That the correct orientation is not known is a sense in which this is not an averaging algorithm.

### 4.3 Asymmetry Detection Algorithm

Since the virus systems we are interested in have asymmetry around 5-fold axis of an icosahedrally symmetric structure, to simplify description, the asymmetry detection algorithm is described with respect to this region. However, the algorithm can be easily generalized to other regions of interest. The overall asymmetric detection algorithm based on MLE as discussed in Section 4.2 is as followed (see Fig. 4.3 & Fig. 4.4 for a detail flow chart): **Step (1)**: Since icosahedrally-symmetric viruses are roughly spherical in shape, we will use a spherically-symmetric reference to identify their locations within a CET tomo-

gram. By hand selection of a small subset of cubes from tomograms and spherically average these cubes as an initial template, we perform template matching analysis to identify virus particle locations within the tomogram. Then, each cube containing one virion is extracted from all tomograms. **Step (2):** Two alternatives: **(a)** Use a maximum likelihood estimator [35, 113] to compute an icosahedrally-symmetric reconstruction of each class of particle from the CET subcubes. This calculation also provides the class label for each CET cube and the orientation of the particle in each CET cube where the orientation is only up to one of the 60 operators from the icosahedral group. **(b)** If only one class of particle is anticipated and if a symmetric reconstruction based on cryo-EM images is available (which presumably has higher resolution than a CET-based reconstruction) then use it as a template to orient the virion in each cube (6 degrees of freedom (DOF), FFT correlation analysis for the 3 translational DOF and exhaustive search for the Euler angles for the 3 rotational DOF where the search tested 5000 orientations which is approximately  $4^\circ$  steps in the angles for this symmetric object). These orientations are only up to an operator from the icosahedral group. **Step (3):** Using the orientations of Step (2), from each cube we extracted 12 subcubes, where each subcube is centered on a 5-fold axis. The extraction is done by rotating the 5-fold axis to the  $+z$  direction using a trilinear interpolator followed by extraction of the samples so that all subcubes are sampled on a cubic lattice where one lattice direction is along the 5-fold axis which also serves to rotate the subcubes to the standard  $+z$  orientation which saves computational effort in later steps. There are 5 different operators from the icosahedral group that can rotate a particular 5-fold to the  $+z$  location and one such operator is arbitrarily chosen. The unresolved 5-fold DOF is determined in the following step. **Step (4):** Sets of twelve subcubes are the input to an improved maximum like-

likelihood classification and reconstruction algorithm, more suitable for processing CET cubes as discussed in Section 4.2. The algorithms work in reciprocal space using a cylindrical harmonic model [91] to describe the subcube. The missing wedge of reciprocal space data is accounted for, specifically, the orientation of the subcube relative to the orientation of the missing wedge is known and the predictions of the reciprocal-space electron scattering intensity are zero in the missing wedge. In this step, if it is known that exactly one vertex deviate from the common symmetric vertex, such as the case of Lambda and P22, such constraint can be enforced in this model approach. Then, the result is a reconstruction for each of the two classes of subcubes, a label which indicates which of the twelve subcubes from one virion is the portal class, and the absolute orientation of each virion (i.e., which of the 60 operators in the icosahedral group will rotate the virion to a standard orientation). **Step (5):** Using the absolute orientation for each particle, rotate the cubes sharing a label to a standard orientation and average over cubes and over symmetries resulting in a reconstruction that will have improved SNR and resolution relative to the original CET cubes and will suffer less or not at all from missing wedge effects. **Step (6):** There are several approaches to exploit the CET results in order to absolutely orient cryo-EM images after they are already oriented up to a rotation from the symmetry group. **(a)** Use correlation or normalized correlation on raw images or difference images (experimental image minus predicted image based on a symmetric reconstruction) with a library of images computed from projections at different orientations of a reconstruction. For difference images, a natural reconstruction is the maximum likelihood reconstruction of the portal from Step (4). For raw images, two natural reconstructions are **(i)** the average of CET cubes aligned by which 5-fold axis is labeled as having the portal by the maximum likelihood

subcube reconstruction algorithm after further averaging over symmetries and **(ii)** insert the maximum likelihood reconstruction of the portal from Step (4) into a symmetric reconstruction created by symmetric processing of the cryo-EM images. Approach (i) has the advantage of the whole reference has uniform resolution. However, CET cubes generally have lower SNR than cryo-EM images and the numbers of CET cubes available are typically smaller than the numbers of cryo-EM images, thus resulting in typically noisy reconstruction. Some form of averaging over symmetries can be applied to the reconstruction to boost SNR in approach (i). Since the reconstruction is only asymmetric at region around only one of the twelve 5-fold axis, but icosahedrally-symmetric everywhere else, the following symmetric averaging scheme is applied: 1) a point and its equivalent positions with respect to icosahedral symmetry are averaged if that point is not within a prescribe region around the unique vertex. 2) intensity inside the unique vertex region is only symmetrically averaged if such symmetry is known (i.e., the unique portal of bacteriophage lambda possess 12-fold symmetry with respect to the 5-fold axis of the particle), otherwise no symmetric averaging is applied. 3) intensities between the icosahedrally averaged capsid and the unique asymmetric region is interpolated based on intensities of the two regions. On the other hand, approach (ii) has the advantage of producing a reconstruction where the symmetric part benefits from the higher spatial resolution of the symmetric cryo-EM reconstruction in comparison with the CET reconstruction. In all three cases, the search is over the 60 different operators of the icosahedral group. **(b)** Our experience is that determination of the correct 5-fold (one of 12) is substantially easier than determination of the correct rotation around that 5-fold (one of 5), partly because reconstruction from CET cubes suffers from lower SNR and smaller data sets. In the case where a

good quality of the unique vertex structure is not possible, we propose to circularly average the unique portal structure with respect to the 5-fold symmetry axis, which improves the SNR, and then any of the approaches of Step (6)(a) can be used to estimate the correct 5-fold. Then, apply the maximum likelihood ideas we have used for an *ab initio* maximum-likelihood reconstruction of infectious bacteriophage P22 [91] to determine which of the 5 rotations. (c) Use the maximum-likelihood estimator of Ref. [91] with the important enhancement of starting from an initial condition computed from any of the templates described in Step (6)(a).

## 4.4 Numerical Results

### 4.4.1 Simulation Study with P22

In order to test the accuracy of the classification algorithm based on MLE, we have performed calculations based on 30 synthetic CET cubes made by deleting the tail of the reconstruction of infectious phage P22 [69], randomly orienting the particle, creating a tilt series of images ( $2^\circ$ ,  $\pm 60^\circ$ ), and using *imod* (<http://bio3d.colorado.edu/imod/>) to create a CET cube. In each simulated CET cube there is one portal 5-fold vertex and eleven non-portal 5-fold vertices. But the fact that each particle has a unique vertex is not used in the calculation. In spite of this, for SNR (squared Euclidean norm of the image divided by the noise variance) greater than 0.1, the calculations result in every particle having a unique vertex which shows the 12-fold symmetry of the portal and not the 5-fold symmetry of the non-portal capsid pentamer. For SNR equal

to 0.01, 28 out of 30 portals are classified as portals and three or four additional non-portal vertices are also classified as portals. But, when normalized cross correlation is computed between the reconstructions and the vertex subcubes of the CET cubes, for particles with two vertices classified as portals, the true portal has a higher correlation value than the incorrect portal. This suggests that if a constraint of only one asymmetric region exists within each particle, the detection rate of portal structures will improve at lower SNR, since the algorithm will pick the one vertex that is most similar to the portal reconstruction (in term of normalized correlation measure) as the portal structure. In the synthetic P22 calculations, we do not see evidence that the locations of vertices with incorrect vertex labels are spatially correlated with the direction of the missing wedge of data.

#### **4.4.2 Lambda Portal Detection and Asymmetric Reconstruction**

Lambda is a bacteriophage consists of an icosahedral capsid and a flexible tail that is connected by a portal to the capsid. Unlike P22 which has a rigid tail, thus making it easy to identify the portal region, there is currently no detail structure of the portal of Lambda, thus making it an excellent example to both test the accuracy of our algorithm on experimental data and applying this novel computational method to answer interesting biological question. Five CET tomograms of purified bacteriophage lambda, each containing  $1024 \times 1024 \times \sim 300$  voxels measuring 7.04 are acquired by collaborators at The Scripps Institute. 210 cubes each measuring  $65 \times 65 \times 65$  voxels and containing one virion are extracted by correlating with a spherically symmetric template. Since the biological specimen being imaged only contains procapsids and a symmetric reconstruction of

the capsid of bacteriophage lambda is available, template matching is used to determine the center location and orientation (only up to one operator of the 60 icosahedral group) of each virion within a CET cube. Using orientation from each cube, we extracted 12 subcubes containing  $27 \times 27 \times 27$  voxels where each subcube is centered on a 5-fold axis and extends from the outer surface of the capsid inward  $27 \times 7.04$ . The subcubes are classified using the maximum likelihood classification and reconstruction algorithm. Since it is known only one portal structure is presented on the procapsid, such constraint is enforced in the model. Furthermore, the common vertices are 5-fold symmetric while the portal structure is known to be a dodecamer [need citations], these symmetries are also enforced in the class reconstruction. This also indicates the advantage of such a model-based approach to reconstruction and classification, in which prior information from other biological experiments can be used favorably to help to improve the accuracy of current calculation. With the class label and the orientation around the 5-fold axis, aligned average is performed for each class of vertices (Fig. 4.1(f-i)). With the location of the portal structure known, absolute orientation of each virion can be determined. Using the absolute orientation of each particle, rotate the cubes sharing a label to a standard orientation and average over cubes (Figure 4.1(b-c)) and over symmetries (Figure 4.1(d-e)) resulting in a reconstruction that will have improved SNR and resolution relative to the original CET cubes and will suffer less or not at all from missing wedge effects. Based on this reconstruction from CET cubes, normalized correlation with raw images are performed to determine absolute projection orientation. With absolute projection orientations, asymmetric reconstruction of Bacteriophage Lambda using filtered backprojection is computed (Figure 4.1(k-m)).



### 4.4.3 STIV Turrets Analysis

STIV are viruses that have turret-like densities at the 5-fold vertexes of its icosahedral capsid, which are often found to be involved in genome translocation [2, 56, 42]. It is known that viruses in the same lineage, such as PRD1, has a specialized 5-fold vertex for DNA packaging, thus, it will be interesting to determine whether this is also true in the case of STIV [42]. To study this, our asymmetry detection algorithm is applied on CET cubes containing STIV particles. Specially, 63 experimental CET cubes, each containing a purified STIV particles, are extracted from five tomograms (Fig. 4.2 (a)). Then, an icosahedral reconstruction and estimated orientation for each particle are computed based on the maximum likelihood approach (Fig. 4.2 (b)). The protruding densities, characteristic of STIV capsid, provide evidence that particle orientation estimation is accurate, otherwise the narrow turret structure will disappear due to inaccurate alignment. Then, turrets structures at 5-fold axes are extracted from each particle as input to the MLE classification algorithm assuming one unique vertex exists per particle. The ML classification and reconstruction results in two somewhat different structures (Fig. 4.2 (c)–(f)). However, no dramatic structural difference are observed between the two classes, as opposed to the strong density that are present interior to the capsid shell for the special vertex but not for the common vertex in the case of bacteriophage lambda. At current resolution, it is uncertain whether the observed minor difference in the two structures is actually biological significant because such a difference might also arise from the difference in SNR between the two reconstructions due to the different numbers of vertex sub-cubes contributing to the reconstructions of the different class. A two class reconstructions without enforcing one special vertex per particle is also computed and the resulting reconstructions between the two classes are

also very similar, leading us to suspect that maybe there is no special vertex that is structurally different that functions as a genome packaging gateway in the case of STIV. However, a more vigorous statistical test and more CET data cubes are required to support such an hypothesis in an convincing way.

#### 4.4.4 Discussion

Though in all the calculation, two classes are assumed in each case, the maximum likelihood estimation approach can also work with multiple classes of vertex, though more classes requires more data, and with constraints on how many instances of a particular class are required in each particle. The issues here modify Steps (4–5) described above. With sufficient data, the number of classes can be large but we have only worked with up to four classes. If any vertex in a particle can have any structure, then the subcubes can be treated as independent, e.g., it does not matter which set of subcubes originated from a single particle. If there are  $N$  particles,  $M$  vertices per particle, and  $J$  classes this implies that the maximum likelihood algorithm works on  $NM$  subcubes and each subcube has a nuisance parameter taking a value in the set  $\{1, \dots, J\}$  which describes its class. This is the situation in the STIV example. Alternatively, if exactly one of the vertices must have a copy of the rare structure then there is dependence among the vertices of each particle. Suppose that there are  $N$  particles,  $M$  vertices, and two classes. Then the maximum likelihood calculation concerns  $N$  pieces of data each of which is  $M$  subcubes and each of which has a nuisance parameter taking a value in the set  $\{1, \dots, M\}$  which describes which of the  $M$  subcubes is the one containing the rare structure. This is the situation in the bacteriophage  $\lambda$  example. If the problem continued to have two classes but

now exactly two vertices had the rare structure, then the nuisance parameter would take a value in the set  $\{1, \dots, M(M-1)\}$  but this and other generalizations have not arisen in Prof. Johnson's virology. Note that the relative locations in the particle of multiple rare structures might be as interesting as the structure of the rare structure and this information can be estimated once the sites are classified.

An advantage of maximum likelihood (ML) is that it hedges, that is, a particular cube or subcube contributes to the reconstruction of each class based on the probability that it belongs to that class. In addition, it contributes at a range of orientations based on the orientation probability distribution that evolves during the reconstruction. Furthermore, when the ML estimator computes class labels, it actually computes the probability mass function on the label values. The natural label estimate is the label with the highest probability, but, if that probability is not sufficiently large, it is possible to discard that data. Similarly, when the ML estimator computes orientations, it actually computes a probability density function (pdf) on the orientation. The natural orientation estimate is the orientation for which the pdf is largest, but if the pdf is not sufficiently narrow, it is possible to discard that data. Further advantages of ML are that it can incorporate the fact that reciprocal-space CET data has a missing wedge and that CET data has low SNR into reconstructions, class labels, and orientations. All the orientations and class labels we propose to use are from ML estimators and therefore have these properties. For reconstructions, we use estimates from both ML (e.g., Figure ??(b)) and an averaging procedure. The averaging procedure is to group all data with the same ML label together, rotate each piece of data by the ML orientation, average the different pieces of data together (e.g., Figures 4.1(b-c)), and finally average over any symmetry that the data should

obey (e.g., Figures 4.1(d–e)). Note that this procedure depends on the ML labels and the ML orientations which are determined jointly from all the data not by a set of pairwise comparisons of two (very noisy) pieces of data as is sometimes done in simpler averaging procedures. The averaging procedure has the disadvantage that accounting for the effect of the missing wedge requires that many different particle orientations are available which is sometimes not true. On the other hand, if a sufficient number of orientations is available, the level of detail shown in the averaging reconstruction reflects the accuracy of the orientations since detail is blurred if the orientations are inaccurate and this is a useful check on the computations.

The overall asymmetric detection algorithm (Fig. 4.3 & Fig. 4.4) might seem unnecessarily complicate at first glance, it actually reflects our efforts to come up with a reliable computational method to deal with the wide range of data that we might have from CET imaging experiments. For example in Step 1 of the algorithm, the option to use ML classification and reconstruction algorithm reflects our anticipation that the CET data cubes might represent a mixture of heterogeneous particles. To ensure that the structural difference that we might detect between ROIs that are related by symmetry within a structure, not structural difference arise from different structural states of a virus, the ML algorithm is used to sort the data set into relative homogeneous subsets that hopefully represent viruses of the same structural state. Similarly, the wide range of options to construct a reference for projection orientation determination of cryo-EM images in Step 6 reflects two challenges we anticipate when dealing with CET data cubes. First, CET data typically has lower SNR and resolution than cryo-EM data, which leads to poorer quality of reconstructions than those from cryo-EM. Thus, different symmetric averaging technique can be applied to improve the

quality of the CET reconstruction, leading to more reliable projection orientation estimation (Step 6 (a), i). Another challenge with reference-based processing is that the result is typically somewhat biased to the reference. Thus, if a poor CET reference is used, it might lead to a wrong reconstruction from cryo-EM data. This leads to the hybrid approach of using both reference-based and reference-free reconstruction algorithms to determine the final asymmetric reconstruction in cryo-EM (Step 6 (b) – (c)). In fact, the overall algorithm presented in Sec. 4 is our effort to utilize both CET and cryo-EM data jointly to determine a high-resolution and asymmetric reconstruction reliably and accurately.

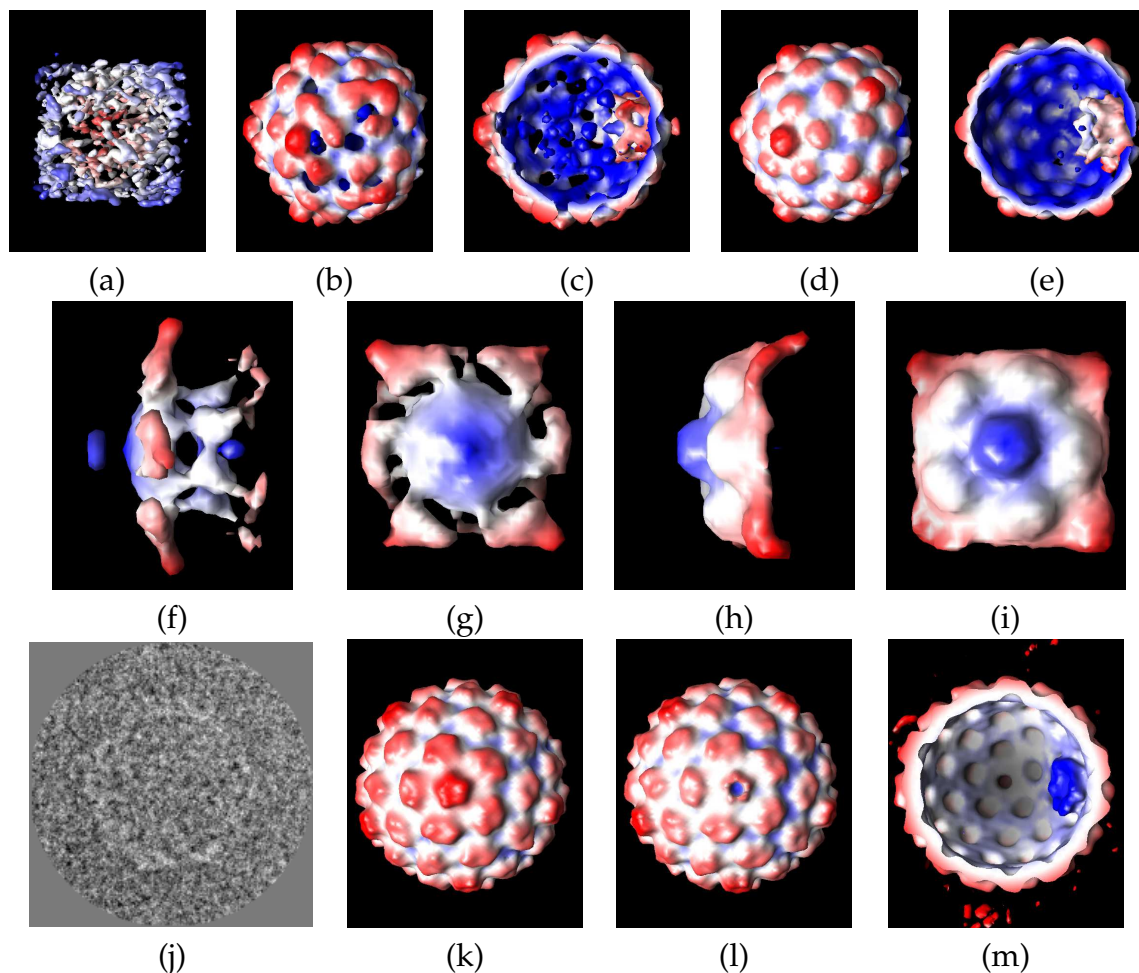


Figure 4.1: Results based on 210 experimental CET cubes and 3500 experimental cryo EM images of bacteriophage  $\lambda$ . Panel (a) Example CET cube. Panels (b-c): Average of CET cubes aligned by which 5-fold axis is labeled as having the portal by the maximum likelihood subcube reconstruction algorithm. Panels (d-e): Average over symmetries of (b-c). Panels (f-g): Portal (6 fold) from (d-e). Panels (h-i): Pentamer (5-fold) from (d-e). Panel (j): Example cryo EM image. Panels (k-l): Asymmetric reconstruction from cryo EM images where (k) and (l) have the common (11 of 12) and the rare (1 of 12) 5-fold axis directed out of the page at the center of the image, respectively, and (m) shows the portal in blue on the right hand side. Different visualization orientations are used in (b-e,m) versus (k-l). (Visualization: UCSF Chimera [89]. Color: distance from origin).

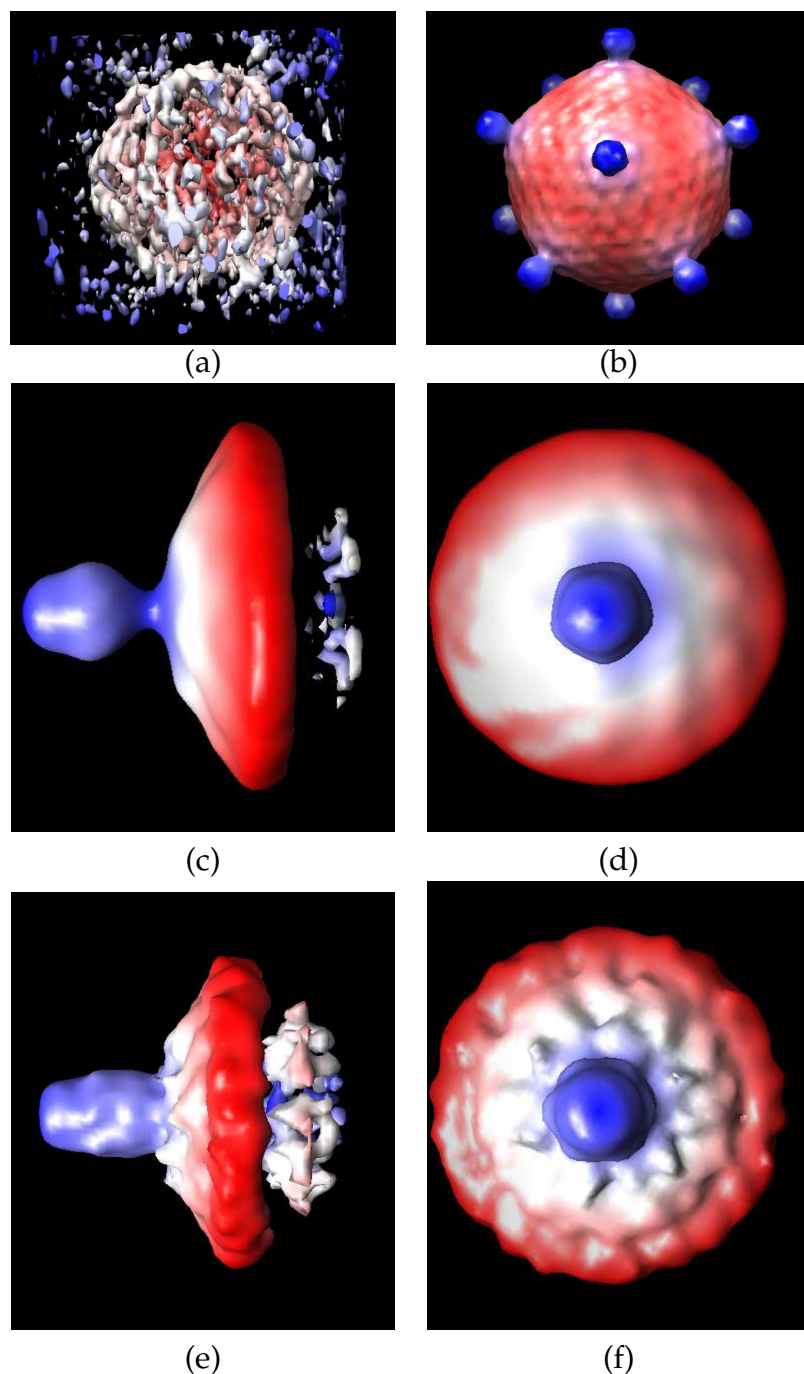


Figure 4.2: Results with 63 experimental CET cubes of purified STIV. Panel (a): Example CET cube. Panel (b): The icosahedrally-symmetric maximum likelihood reconstruction [Step (2)(a)]. Panels (c–f): Reconstructions of Class 1 [(c–d)] and Class 2 [(e–f)]. Each subcube receives a class label and a rotational orientation [Step (4)]. After rotation to an absolute orientation, the subcubes sharing a class label are averaged and the result is further averaged over the rotational symmetry [5-fold for (c–d) and 6-fold for (e–f)] to compute (c–f). (Visualization: UCSF Chimera [89]. Color: distance from origin).

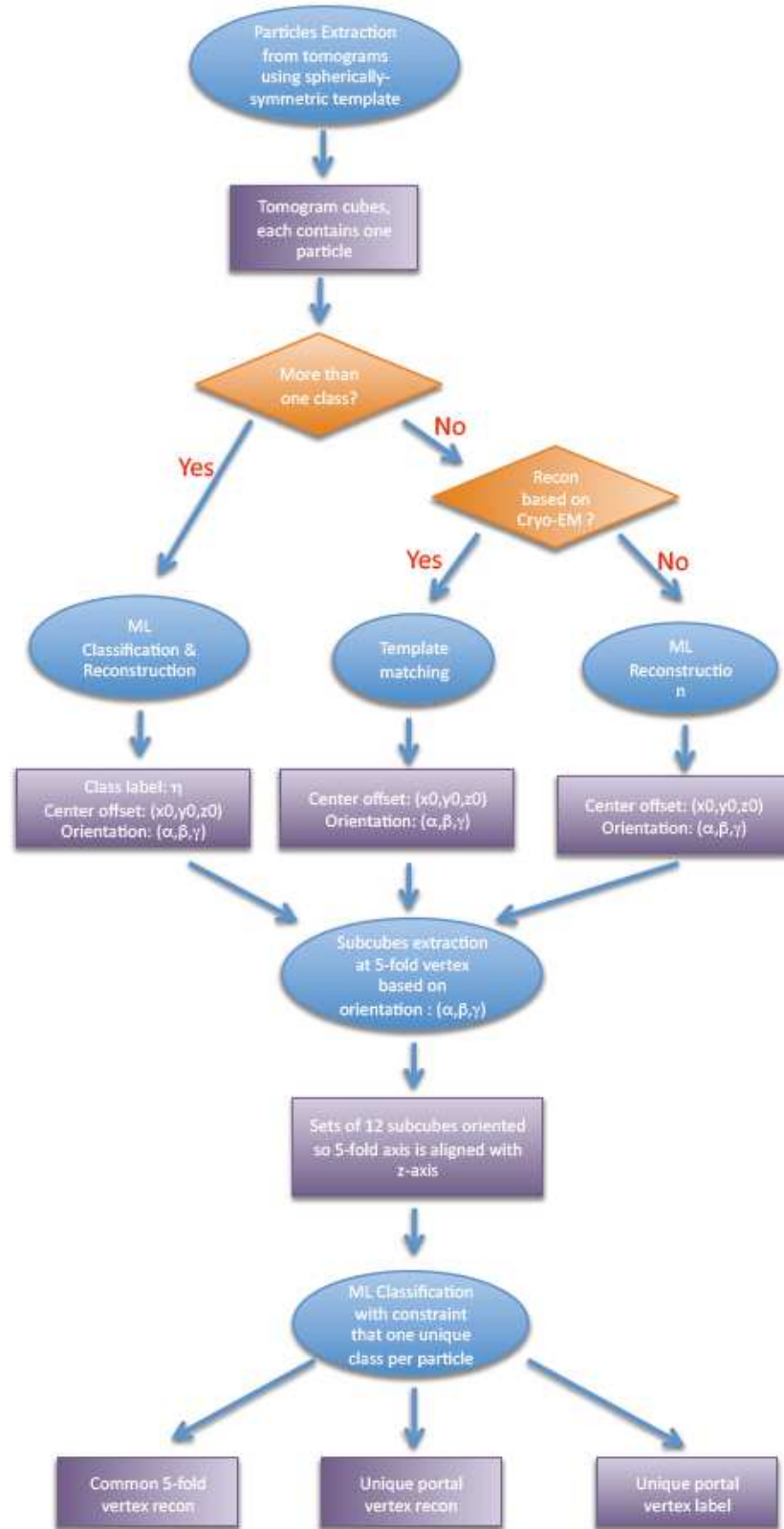


Figure 4.3: Asymmetric Detection Algorithm Float Chart Part (A)



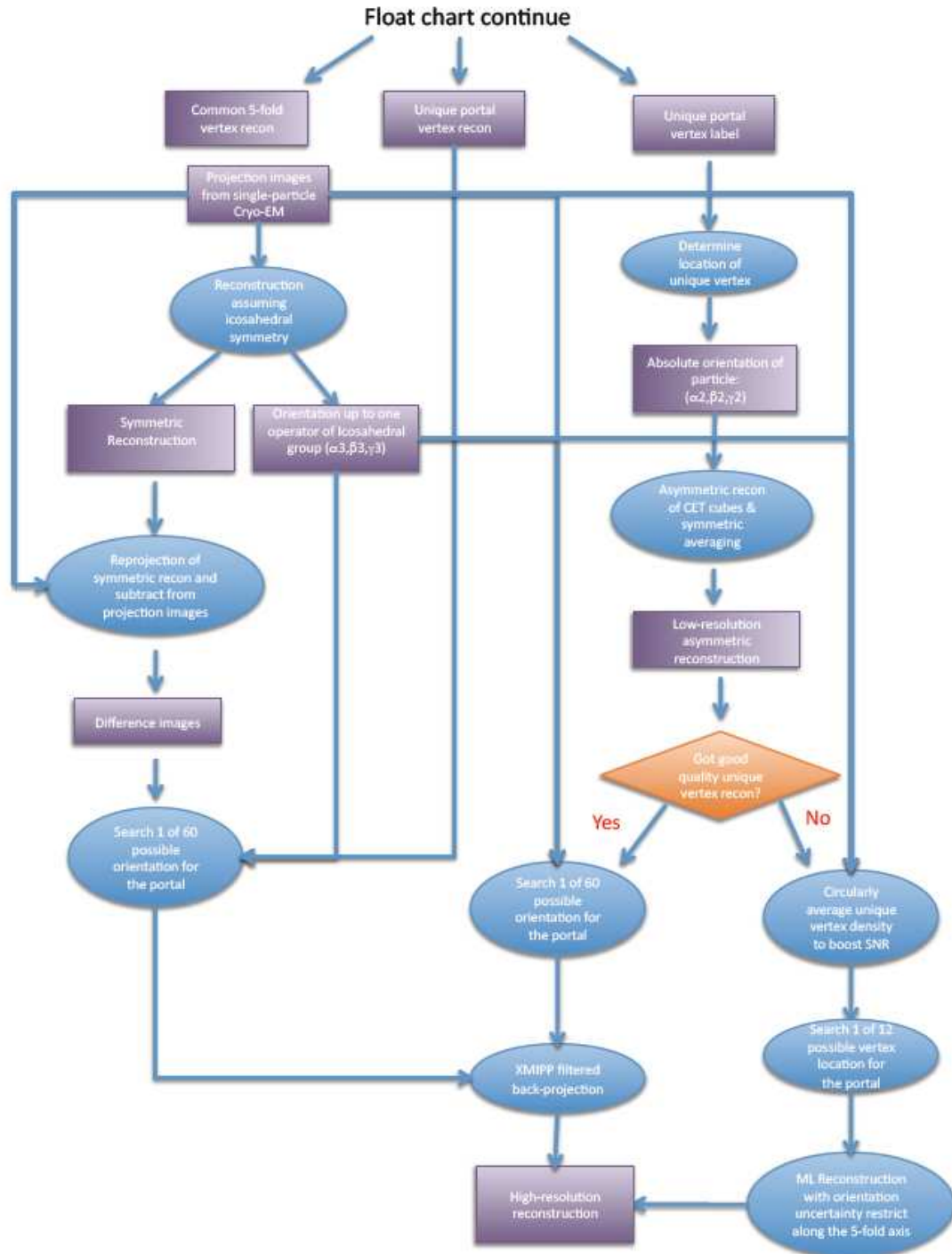


Figure 4.4: Asymmetric Detection Algorithm Float Chart Part (B)

CHAPTER 5

EFFICIENT COMPUTATION OF MAXIMUM LIKELIHOOD ESTIMATOR  
(MLE) VIA SPHERICAL FOURIER TRANSFORM

## 5.1 Introduction

In Chapter 2, we introduce a maximum likelihood (ML) algorithm to determine a set of distinct structures from a set of heterogeneous noisy CET data cubes, each containing a particle at an unknown orientation and each having a region of missing data in Fourier space. We have demonstrated that the ML approach is robust against noise because the algorithm does not deterministically assign a class membership and particle orientation to each data cube as is done in more traditional method [31, 100, 8]. Instead, a complete statistical description of how a data cube belongs to a specific class or takes on a specific orientation is computed. Then, this quantitative description of uncertainty is incorporated into the determination of the underlying structures that give rise to the experimental data cubes in such a way that if a data cube has high uncertainty in its orientation or its class membership, it contributes less to the determination of the structure of that class.

However, such a detail statistical description makes the computation intensive, which is unappealing relative to some of the more deterministic approaches. Specifically, in order to capture the uncertainty in a particle's orientation in a CET cube, a prediction based on an estimate of the underlying structure is compared with each CET data cube at each orientation. This requires the computation of a mathematical model of the underlying structure at all possible orientations sampled on a cartesian grid since CET data cubes are

recorded naturally on such a grid, which is a computationally intensive task.

Recently, Kostelec et. al. developed an efficient way to compute correlation between two signals defined on a sphere by first transforming the signals into coefficients of spherical harmonic series and then utilizing the Fourier Transform on a rotational group or SO(3) Fourier Transform (SOFT) to arrive at the answer quickly. [47, 63]. We generalize such ideas to 3-D signals, in which we first obtain the Spherical Fourier Transform (SFT) coefficients of a 3-D signal and then the correlation between a signal and a rotated version of another can be computed efficiently using SOFT. In this Chapter, we show how one can compute the MLE more efficiently utilizing these transformation techniques. The detailed formulation of the Spherical Fourier Transform and SO(3) Fourier Transform is discussed in Appendix B.

## 5.2 Maximum Likelihood Formulation of the Cryo-ET Problem

In the Cryo-ET problem, we have a set of data heterogeneous CET cubes, each containing a particle of interest extracted from a tomogram. The goal is to determine the set of distinct 3D structures giving rise to the set of heterogeneous experimental data cubes. A data cube  $Y(k)$  and its underlying 3D structure in Fourier space is related by the following statistical relationship:

$$Y(k) = W(k) \cdot \Lambda(g)X(k) + V(k) \quad (5.1)$$

where (1)  $\cdot$  is point-wise multiplication. (2)  $\Lambda(\cdot)$  represents the rotational operation and  $g \in \text{SO}(3)$  represents a specific rotation within the rotation group SO(3) that describes how the underlying structure is oriented in a CET cube. (3)  $W(k)$  is a binary mask function that models the missing region of Fourier space

in each reconstructed data cube due to incomplete projection information, characteristic of Cryo-ET reconstruction. (4)  $V(k)$  represents the zero-mean white additive Gaussian noise with power spectral density  $\sigma^2$  that represent noise arising from the imaging experiment. The above mathematical model relating  $Y(k)$  and  $X(k)$  implies a conditional likelihood function of the following form given a specific  $g$  and  $X$ :

$$p(Y|g, X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\|Y - W \cdot \Lambda(g)X\|^2}{2\sigma^2}} \quad (5.2)$$

However, another characteristic of the Cryo-ET problem is that the orientation of a particle inside the tomogram is unknown (i.e.,  $g$  is unknown). In these situations, the unknown parameters are marginalized in order to obtain an unconditional likelihood function  $p(Y|X)$  that relates just the observation ( $Y$ ) with the desired unknown parameters ( $X$ ):

$$P(Y|X) = \int_{g \in SO(3)} p(Y|g, X) dg \quad (5.3)$$

Then, in the maximum likelihood approach, the best estimate of the underlying structure,  $\hat{X}$ , is one that maximize the unconditional likelihood function  $p(Y|X)$ .

There are many ways to represent rotations. One way to represent all possible rotations within the rotational group (  $SO(3)$  ) is by a set of Euler angles  $SO(3) = \{(\alpha, \beta, \gamma) | 0 \leq \alpha \leq 2\pi, 0 \leq \beta \leq \pi, 0 \leq \gamma \leq 2\pi\}$ . Integrating over all possible rotations in Eq. 5.3 means a multi-dimensional integral over the three Euler angles  $(\alpha, \beta, \gamma)$  which means evaluating the conditional likelihood function  $p(Y|g, X)$  at many different  $g$ . Much of the computation is in computing the quadratic term  $\|Y - W \cdot \Lambda(g)X\|$  in the conditional likelihood function. The quadratic term of the conditional likelihood function, denoted  $q(g)$ , can be expanded in the following way:

$$q(g) = \|Y(k) - W(k) \cdot \Lambda(g)X(k)\|^2$$

$$\begin{aligned}
&= \|Y(k)\|^2 + \|(W(k) \cdot \Lambda(g)X(k))\|^2 - 2 \langle Y(k), W(k) \cdot \Lambda(g)X(k) \rangle \\
&= \|Y(k)\|^2 + \langle W(k) \cdot W(k), \Lambda(g)(X(k) \cdot X(k)) \rangle \\
&\quad - 2 \langle W(k) \cdot Y(k), \Lambda(g)X(k) \rangle
\end{aligned} \tag{5.4}$$

The last equality holds because the inner product of three functions are invariant with respect to the order of how the functions are multiplied together. The second and third term of Eq. 5.4 can be viewed as correlation between one signal and a rotated version of a second one, which can be computed efficiently using SOFT after obtaining the SFT coefficients of each 3-D signal (see Appendix B.2).

Another thing to notice is that, we can multiply both signals inside the quadratic term by a rotation operation, since rotation is unitary. Specifically, if we apply a rotation that is the inverse of the rotation  $\Lambda(g)$ , denoted  $\Lambda^{-1}(g)$ , we can eliminate the rotation operator applied on the underlying structure.

$$\|Y - W \cdot \Lambda(g)X\|^2 = \|\Lambda^{-1}(g)(Y - W \cdot \Lambda(g)X)\|^2 \tag{5.5}$$

$$= \|\Lambda^{-1}(g)Y - (\Lambda^{-1}(g)W) \cdot X\|^2 \tag{5.6}$$

Formulated this way, we will need to rotate the experimental data cube  $Y(k)$  into different orientations rather than the underlying structure we desire. Rotating data sampled on a Cartesian grid requires interpolation between sampled points, which is computational costly and prone to error when such operation is performed on a noisy data cube, especially a cube in reciprocal space. Thus, we favor the formulation  $\|Y - W \cdot \Lambda(g)X\|$  in Chapter 2, since rotation operation is exact on the mathematical model  $X(k)$ . In this new approach, we propose to resample the experimental data cubes in spherical coordinates and analyze the data after SFT transform, in which rotation operation is exact and fast (see

Appendix B.1.2).

Furthermore, the new quadratic form implies a much simpler form of the solution in the Maximization step of the Expectation-Maximization (EM) algorithm used to solve the optimization problem of finding a  $\hat{X}$  that maximizes the likelihood function  $p(Y|X)$ . In each step of the EM algorithm, we construct a lower bound for the unconditional likelihood function  $p(Y|X)$  based on a previous estimate, denoted  $X^{n-1}$ . The lower bound, denoted  $Q(X|X^{n-1})$ , in terms of Eq. 5.6 and Eq. 5.6 is as follows:

$$Q(X|X^{n-1}) = \int_{g \in SO(3)} \|Y - W(k) \cdot \Lambda^{-1}(g)X\|^2 p(g|Y, X^{n-1}) dg \quad (5.7)$$

$$= \int_{g \in SO(3)} \|\Lambda^{-1}(g)Y - \Lambda^{-1}(g)W(k) \cdot X\|^2 p(g|Y, X^{n-1}) dg \quad (5.8)$$

Then, the new estimate of  $X$ , denoted  $X^n$ , is the one that maximizes the lower bound function  $Q(\cdot|X^{n-1})$ . Since  $Q(X|X^{n-1})$  is the lower bound for the likelihood function  $p(Y|X)$ , the new estimate  $X^n$  will always have greater likelihood value than the old estimate  $X^{n-1}$ , thus we always obtain an estimate of  $X$  that increases the likelihood function in each step and eventually we will reach a local maximum.

Because integration is a linear operation, Eq. 5.7 and Eq. 5.8 implies a least squares problem in term of the unknown  $X$ . Thus, the maximum is the solution to the following linear system,

$$\left( \int_{g \in SO(3)} \Lambda(g)W(k)p(g|Y, X^{n-1})dg \right) \cdot X = \int_{g \in SO(3)} \Lambda(g)Yp(g|Y, X^{n-1})dg \quad (5.9)$$

which can be solved by simple element-wise division because the left hand side of Eq 5.9 is simply a element-wise product of two functions.

$$X = \frac{\int_{g \in SO(3)} \Lambda^{-1}(g)Yp(g|Y, X^{n-1})dg}{\int_{g \in SO(3)} \Lambda^{-1}(g)Wp(g|Y, X^{n-1})dg} \quad (5.10)$$

However, such formulation is not possible for Eq. 5.7 because the rotation operation  $\Lambda(g)$ , is a nontrivial linear operation on  $X$ . Eq. 5.10 implies that instead of solving a large linear system in each step of the EM algorithm which requires  $O(N^2)$ , where  $N$  is the number of voxels in a data cube, the solution is a simple weighted averaging, which requires only  $O(N)$  operations.

This new statistical formulation and the advantage of SFT implies the following more efficient algorithm to calculate the MLE: 1) Resample data cubes in spherical coordinates and perform SFT. 2) Initialize a guess for  $X$ , denoted  $X^{n-1}$  and transform this to SFT coefficients. 3) Compute the conditional likelihood function ( $p(Y|g, X^{n-1})$ ) efficiently utilizing SOFT, and based on the conditional likelihood function, compute the posterior probability density (pdf) function  $p(g|Y, X^{n-1})$  using Baye's rule. 4) Rotate each data cube to each orientation by a linear transform of the coefficients from SFT 5) Sum over all possible rotated coefficients of each data cube and over all data cubes with weight  $p(g|Y, X^{n-1})$ . 6) Keep track of the weights contributes to each voxel and sum over these weights 7) Transform the weighted sum of SFT coefficients back to signal in standard Fourier space and divide by the weight contribute to each voxel to obtain the correct normalization. 8) This weighted average will be our new estimate of  $X$ , thus repeat Step (2) – Step(7) with this new estimate until the old and new estimate converges.

### 5.3 Practical Implementation

In order to obtain the coefficients for SFT, we need to take the inner product between a 3-D signal and each basis function. In actual implementation, this

implies numerical integration of the product of the function sampled in spherical coordinates and each basis function. However, experimental data cubes are naturally sampled on a cartesian grid of  $N$  sample points in each direction. To sample sufficiently in spherical coordinates to cover the whole cartesian cube, we need a sphere that circumscribes the cartesian grid. This requires the number of radial sampled points,  $N_r$ , to be at least  $\sqrt{3}\frac{N}{2}$ , and the number of sampling points in both angular directions,  $N_\theta$  and  $N_\phi$ , to be at least  $\sqrt{3\pi}N$ . To perform re-sampling on the spherical coordinates, trilinear interpolation between the cartesian grid points are used in the current implementation.

In evaluating the inner product between the spherical sampled data cubes, and each basis function, we first perform the spherical harmonic (SH) expansion between the signal at each radius, since there is FFT-like algorithm that perform the SH transformation efficiently and reliably [47]. Next, we perform a numerical spherical bessel transform on SH coefficients of the same order of  $(l,n)$  along each radius using simple trapezoid rules since the data cubes are sampled regularly along the radial direction. From our experience, if we sample the basis functions at greater than fifty points, the numerical integration between pair of radial basis functions give a reasonable orthonormality result. Thus, in the implementation, we will sample data cubes along the radial direction at least 50 points when  $N_r$  is smaller than this.

After obtaining the coefficients of SFT using numerical integration, most of the computations involving rotations can be computed efficiently. In this new formulation, the most time-consuming part is evaluating Eq. 5.10. In practical implementations, we are again faced with the following two numerical integrations, where we explicitly express a rotation ( $g$ ) as three Euler angles ( $\alpha, \beta, \gamma$ ) and



$w(\cdot)$  are the weights used in the numerical integration for each variable:

$$X = \frac{\sum_{\alpha} \sum_{\beta} \sum_{\gamma} \Lambda(\alpha, \beta, \gamma) Y p(\alpha, \beta, \gamma | Y, X^{n-1}) w(\alpha) w(\beta) w(\gamma)}{\sum_{\alpha} \sum_{\beta} \sum_{\gamma} \Lambda(\alpha, \beta, \gamma) W p(\alpha, \beta, \gamma | Y, X^{n-1}) w(\alpha) w(\beta) w(\gamma)} \quad (5.11)$$

Since the Euler angles must be uniformly sampled in order to utilize the SOFT transform, the weights correspond to simple trapezoid rule in each variable.

This triple sum is computationally expensive. For example, if we sample every  $5^\circ$  in each Euler angle, the total number of SFT coefficients to be rotated and summed over will be greater than  $10^6$ . Typically, the posterior probability  $p(\alpha, \beta, \gamma | Y, X^{n-1})$  falls off sharply away from the location of maximum value, thus, we can approximate the new estimate by truncating the sum to a subset such that the weight (i.e., posterior probability  $p(\alpha, \beta, \gamma | Y, X^{n-1})$ ) is greater than a certain threshold to accelerate the computation.

A parallel version of the algorithm is also implemented using the MPI library. According to Eq. 5.11, to obtain a new estimate of  $X$  from the previous step in the EM algorithm, we need to perform the exact same operation (rotate its SFT coefficients and weighted sum all coefficients) on each data cube, thus leading to a simple implementation of single process multiple data (spmd) parallelism, in which each process is responsible for performing the operations on a subset of the data.

## CHAPTER 6

# UNDERSTANDING CONFORMATIONAL DYNAMICS OF MACROMOLECULAR COMPLEXES FROM THE HETEROGENEITY OF CRYO-EM DATA

### 6.1 Introduction

The traditional paradigm in studying macromolecular assemblies is that sequence encodes 3D structure, and 3D structure in turn determines function of the molecule. However, advances in structural genomics have shown that conformation alone cannot explain the mechanism of biological function, but rather the conformational dynamics of the macromolecular assemblies are crucial elements to explain its function [92]. For example, a recent study on a range of 98 enzymes structures shows that the catalytic site is always at or near hinges of global motion of protein domains no matter what their enzymatic activities are, indicating that mechanical and chemical properties are coupled together to facilitate catalytic activities of enzymes [117]. Thus, it is important to study the conformational dynamics of a macromolecule as well as their 3-D conformational state.

Many physical models have been proposed to study the dynamics of protein structures. Molecular dynamics (MD) and Gaussian Network Model (GNM) probably represent two ends of the spectrum. In MD, interaction between each pair of amino acid residues and each type of interaction is accounted for. Thus, the computational complexity is huge in such method. At the other end of the spectrum, GNM models are a coarse-grained model that is motivated by the physics of polymer networks, in which point masses representing blocks of

residues are connected by linear springs [110].

In such a model, the object is described as a collection of "springs and masses" which obey classical mechanics. Using standard ideas of statistical mechanics, the appropriate ensemble to describe the population of objects in equilibrium with a heat bath is the canonical ensemble. From the statistic mechanics, the mean square fluctuation of each residue can be computed. In previous studies, the underlying mechanical parameters (i.e., spring constants of the model) are estimated by minimizing the difference between the predicted mean square fluctuations with experimental fluctuations measured by x-ray crystallography (e.g., the B-factor). With such a mechanical model, useful properties of the structural dynamics of the macromolecular assemblies can be extracted by performing normal model analysis (NMA). To this end, the GNM model is attractive, since it has been shown that together with NMA analysis, this model is able to predict large domain motions of structures as accurately as those using more detailed model (such as MD) [92].

In GNM, model parameters are usually estimated based on experimental data. Typically, a single parameter is assumed for all spring constants and is obtained by enforcing the average B-factor values between experimental and theoretical values. However, the residue-specific agreement between the two is not always good. Erman extended the model by using a multi-parameter GNM model to make the theoretical and experimental B-factors agree on a per residue level, which is useful in the study of drug binding [29]. However, the fact that crystallographic data are needed to estimate the parameters of the underlying model impairs its usefulness in the study of conformational dynamics of large macromolecular assemblies which are simply too large to be crystallized.

In single-particle cryo-EM or CET data, the specimen, which is at equilibrium at roughly room temperature, is flash frozen to liquid nitrogen temperatures or lower before being placed in the electron microscope for imaging. Any motion of the object that has a time constant which is large compared to the freezing time is trapped. These motions typically are concerted motions of large domains of a macromolecular assembly that are transient. Since the freezing time is short, the goal is to freeze so quickly that water cannot form crystalline ice but rather forms vitreous ice, and the objects are large on the molecular scale, e.g., the capsid of the bacteriophage Hong Kong 97 measures roughly  $650\text{\AA}$  in diameter, significant dynamics may be trapped by the freezing, thus, multiple images shown multiple instances of the temporal-dynamics of the basic structure. Therefore, structural dynamics contributes a significant part to the variability of the voxel values of the reconstruction. In statistical sense, this results in a nonzero covariance of the voxel values.

Currently, two different approaches have been developed to estimate the covariance of the reconstruction maps from single-particle cryo-EM data. The first one is an extension to the maximum-likelihood approach presented in [24], in which joint estimation of mean and covariance are performed on the available data set [125]. Another way is using a resampling method such as bootstrapping, in which many reconstructions are calculated based on random subsets of the original image data, then the covariance is estimated from this ensemble of reconstructions [120]. In both cases, true noise unrelated to the structure is estimated from background pixels of the image data in order to decouple the variability induced from structural dynamics [125, 120].

From the mechanical model, theoretical mean and covariance of the 3-D elec-

tron scattering intensity of each voxel of a structure can be computed and the results are symbolic formulas. Therefore there are two sets of mean and covariance, one from the data and one from the mechanical model. By comparing these two sets, the parameters of the mechanical model can be determined. More specifically, the model parameters can be estimated in two steps: (1) Estimate the mean and the covariance from the data using one of the two methods mentioned above. (2) Fit the parameters of the symbolic formulas to the estimates from the data. Essentially, the variance of cryo-EM data provides yet another source of experimental information to estimate the model parameters for this useful, albeit coarse-grained, mechanical model. This opens up the possibility of applying the GNM model in the study of structural dynamics of a much broader class of macromolecules.

## 6.2 Mechanical model

The mechanical model is a collection of point masses that interact via a potential energy function that is a function of position only. The  $N_p$  masses are enumerated by  $\alpha \in \{1, \dots, N_p\}$ . The equilibrium positions of the masses are  $x_{0,\alpha} \in \mathbb{R}^3$  and the positions of the masses are  $x_\alpha = x_{0,\alpha} + \delta_\alpha$  where  $x_\alpha \in \mathbb{R}^3$  and  $\delta_\alpha \in \mathbb{R}^3$ . The equilibrium is assumed to be stable. Let the time derivative of  $\delta_\alpha$  be denoted by  $\dot{\delta}_\alpha$ . Let  $\boldsymbol{\delta} = (\delta_1^T, \dots, \delta_{N_p}^T)^T$ ,  $\dot{\boldsymbol{\delta}} = (\dot{\delta}_1^T, \dots, \dot{\delta}_{N_p}^T)^T$ ,  $\mathbf{x} = (x_1^T, \dots, x_{N_p}^T)^T$ , and  $\mathbf{x}_0 = (x_{0,1}^T, \dots, x_{0,N_p}^T)^T$ .

Let  $m_\alpha > 0$  be the mass of the  $\alpha$ th mass. The kinetic energy of the system of

masses is denoted by  $\mathcal{T}(\dot{\delta})$  and defined by

$$\mathcal{T}(\dot{\delta}) = \sum_{\alpha=1}^{N_p} \frac{1}{2} m_{\alpha} \left\| \frac{d(x_{0,\alpha} + \delta_{\alpha})}{dt} \right\|^2 = \sum_{\alpha=1}^{N_p} \frac{1}{2} m_{\alpha} \|\dot{\delta}_{\alpha}\|^2 = \frac{1}{2} \dot{\delta}^T T \dot{\delta} \quad (6.1)$$

where  $T \in \mathbf{R}^{3N_p \times 3N_p}$  is a diagonal matrix where  $T_{3\alpha-2,3\alpha-2} = T_{3\alpha-1,3\alpha-1} = T_{3\alpha,3\alpha} = m_{\alpha}$ . Therefore,  $T$  is a real-valued Hermitian-symmetric positive-definite matrix.

The potential energy of the system of masses is denoted by  $\mathcal{V}_{\mathbf{x}_0}(\delta)$ . Only small perturbations around the equilibrium position  $\mathbf{x}_0$  are considered, i.e.,  $\|\delta\|$  is small. Consider the Taylor series of  $\mathcal{V}_{\mathbf{x}_0}(\delta)$  with respect to  $\delta$  around  $\delta = \mathbf{0}_{3N_p}$ . By redefining the zero of potential energy, it can be assumed without loss of generality that the constant term, i.e.,  $\mathcal{V}_{\mathbf{x}_0}(\mathbf{0}_{3N_p})$ , satisfies  $\mathcal{V}_{\mathbf{x}_0}(\mathbf{0}_{3N_p}) = 0$ . The coefficients of the linear term are the partial derivatives

$$\left. \frac{\partial \mathcal{V}_{\mathbf{x}_0}(\delta)}{\partial (\delta_{\alpha})_i} \right|_{\delta=\mathbf{0}_{3N_p}} \quad (6.2)$$

for  $\alpha \in \{1, \dots, N_p\}$  and  $i \in \{1, 2, 3\}$ . Because  $\mathbf{x}_0$  is an equilibrium, these coefficients are all 0. The coefficients of the quadratic term are the partial derivatives

$$V_{\alpha,i;\alpha',i'} = \left. \frac{\partial^2 \mathcal{V}_{\mathbf{x}_0}(\delta)}{\partial (\delta_{\alpha})_i \partial (\delta_{\alpha'})_{i'}} \right|_{\delta=\mathbf{0}_{3N_p}} \quad (6.3)$$

for  $\alpha, \alpha' \in \{1, \dots, N_p\}$  and  $i, i' \in \{1, 2, 3\}$ . Because  $\mathcal{V}_{\mathbf{x}_0}(\cdot)$  is real valued, it follows that  $V_{\alpha,i;\alpha',i'} \in \mathbf{R}$ . Because the order of the partial derivatives does not matter, it follows that  $V_{\alpha,i;\alpha',i'} = V_{\alpha',i';\alpha,i}$ . Because the equilibrium is stable, it follows that the matrix  $V_{\mathbf{x}_0} \in \mathbf{R}^{3N_p \times 3N_p}$  with elements  $V_{\alpha,i;\alpha',i'}$  is positive semidefinite. Taking the first nonzero term in the Taylor series for small  $\|\delta\|$ , it follows that  $\mathcal{V}_{\mathbf{x}_0}(\delta)$  can be approximated by

$$\mathcal{V}_{\mathbf{x}_0}(\delta) = \frac{1}{2} \delta^T V \delta \quad (6.4)$$

where  $V \in \mathbf{R}^{3N_p \times 3N_p}$  is a real-valued Hermitian-symmetric non-negative matrix. A natural potential energy function is implied by the following assumptions.

Suppose that the  $\alpha$ th mass is bound to neighbors  $\alpha' \in \mathcal{N}_\alpha$  by linear springs with spring constants  $k_{\alpha,\alpha'}$  and rest lengths  $b_{\alpha,\alpha'}$ . Then the potential energy function is

$$\mathcal{V}_{\mathbf{x}_0}(\boldsymbol{\delta}) = \sum_{\alpha=1}^{N_p} \sum_{\alpha' \in \mathcal{N}_\alpha} \frac{1}{2} k_{\alpha,\alpha'} \left[ \left\| (x_{0,\alpha} + \delta_\alpha) - (x_{0,\alpha'} + \delta_{\alpha'}) \right\| - b_{\alpha,\alpha'} \right]^2. \quad (6.5)$$

The total energy of the system of masses is denoted by  $\mathcal{E}_{\mathbf{x}_0}(\boldsymbol{\delta}, \dot{\boldsymbol{\delta}})$  and defined by

$$\mathcal{E}_{\mathbf{x}_0}(\boldsymbol{\delta}, \dot{\boldsymbol{\delta}}) = \mathcal{T}(\dot{\boldsymbol{\delta}}) + \mathcal{V}_{\mathbf{x}_0}(\boldsymbol{\delta}). \quad (6.6)$$

The model described in this section has a finite number of point masses and therefore a finite number of degrees of freedom. Having a finite number of degrees of freedom simplifies the mathematics. However, unless the data supports atomic resolution, a continuum model is more natural. While all of the same ideas can be applied to a continuum model, the fact that the model has an infinite number of degrees of freedom complicates the mathematics and so none of that mathematics is shown. In Section 6.3 a smoothing parameter denoted by  $\delta$  is introduced into the point mass model and the smoothed point mass model is a model with a finite number of degrees of freedom and smooth behavior so it may represent a good compromise between point mass and continuum models.

### 6.3 Electron scattering intensity model

The basic model is that the  $\alpha$ th mass is an impulsive electron scatterer with weight  $f_\alpha$ : Let  $x \in \mathbb{R}^3$ . Then the impulsive scatterer model is

$$\rho(x) = \sum_{\alpha=1}^{N_p} f_\alpha \delta(x - (x_{0,\alpha} + \delta_\alpha)). \quad (6.7)$$

Because of the impulsive nature of the left hand side of Eq. 6.7, it is not possible to multiply  $\rho(x)$  by  $\rho(x')$  which is necessary in order to define the auto-correlation function of  $\rho(\cdot)$ . Therefore, define a smoothed electron scattering intensity, denoted by  $\rho_q(\cdot)$ , by

$$\rho_q(x) = \rho(x) * N(\mathbf{0}_3, q^2 I_3)(x) \quad (6.8)$$

where  $*$  is 3-D convolution and where it is natural to use a covariance matrix, i.e.,  $q^2 I_3$ , that is proportional to the identity matrix, i.e., isotropic, because the model is a point mass model. Expanding Eq. 6.8 gives

$$\rho_q(x) = \sum_{\alpha=1}^{N_p} f_{\alpha} N(\mathbf{0}_3, q^2 I_3)(x - (x_{0,\alpha} + \delta_{\alpha})). \quad (6.9)$$

Let  $E$  denote expectation where the expectation is taken over the positions of the masses, i.e.,  $\delta$ . Let the mean and auto correlation functions of  $\rho_q(\cdot)$  be denoted by  $\bar{\rho}_q(x)$  and  $R_{\rho_q, \rho_q}(x_a, x_b)$ , respectively, and defined by

$$\bar{\rho}_q(x) = E[\rho_q(x_a)] = \sum_{\alpha=1}^{N_p} f_{\alpha} E \left[ N(\mathbf{0}_3, q^2 I_3)(x - (x_{0,\alpha} + \delta_{\alpha})) \right] \quad (6.10)$$

and

$$R_{\rho_q, \rho_q}(x_a, x_b) = E[\rho_q(x_a) \rho_q(x_b)] \quad (6.11)$$

$$= \sum_{\alpha=1}^{N_p} \sum_{\alpha'=1}^{N_p} f_{\alpha} f_{\alpha'} E \left[ N(\mathbf{0}_3, q^2 I_3)(x_a - (x_{0,\alpha} + \delta_{\alpha})) N(\mathbf{0}_3, q^2 I_3)(x_b - (x_{0,\alpha'} + \delta_{\alpha'})) \right],$$

respectively, and let the covariance function of  $\rho_q(\cdot)$  be denoted by  $C_{\rho_q, \rho_q}(x_a, x_b)$  and defined by

$$C_{\rho_q, \rho_q}(x_a, x_b) = E[(\rho_q(x_a) - \bar{\rho}_q(x_a))(\rho_q(x_b) - \bar{\rho}_q(x_b))] = R_{\rho_q, \rho_q}(x_a, x_b) - \bar{\rho}_q(x_a) \bar{\rho}_q(x_b). \quad (6.12)$$



## 6.4 Statistical mechanics

The system is treated as a classical system. The particles are distinguishable, both by different equilibrium locations and by potentially different masses. Before the snap-freezing event, the objects are at equilibrium at constant temperature  $T$  which is near room temperature. Therefore the appropriate statistical mechanical ensemble is the canonical ensemble in which the system of interest is in equilibrium with a heat bath. The freezing is rapid since it is necessary to prevent the formation of crystalline ice in the specimen and, instead, vitreous ice is formed. Any type of motion of the object with a longer characteristic time scale than the freezing time scale would be trapped. In the canonical ensemble, the microscopic degrees of freedom take all possible values and the probability of taking a particular value is proportional to  $\exp(-\beta E)$  where  $\beta = 1/(kT)$  and  $E$  is the total energy [49, p. 144, Eq. 7.5]. In this system, the microscopic degrees of freedom are  $(\delta^T, \dot{\delta}^T)^T \in \mathbb{R}^{6N_p}$  and, in particular, all values in  $\mathbb{R}^{6N_p}$  are permitted. In this system, the energy is the quadratic form in  $(\delta^T, \dot{\delta}^T)^T$  described by Eqs. 6.1, 6.4, and 6.6. Because all values in  $\mathbb{R}^{6N_p}$  are permitted and the energy is quadratic in the microscopic degrees of freedom, the pdf for the microscopic degrees of freedom is a Gaussian pdf with mean and covariance denoted by  $m$  and  $Q$ , respectively, and defined by

$$m = \mathbf{0}_{6N_p} \quad (6.13)$$

$$Q = \beta \begin{bmatrix} V^{-1} & \mathbf{0}_{3N_p, 3N_p} \\ \mathbf{0}_{3N_p, 3N_p} & T^{-1} \end{bmatrix}. \quad (6.14)$$

## 6.5 Implications of Sections 6.2–6.4

**Introduction** First, note that since the pdf on the microscopic degrees of freedom is Gaussian, all of the marginal pdfs are also Gaussian with means and covariances that can be read off from  $m$  and  $Q$ . Second, note that Eqs. 6.10 and 6.12 involve  $\delta$  but not  $\dot{\delta}$ . Therefore all the results are determined by the marginal pdf on  $\delta$  and this pdf does not involve the matrix  $T$  that describes the kinetic energy of the system, e.g., does not involve the masses of the point masses.

Define  $V_{\alpha,\alpha'}^{-1} \in \mathbb{R}^{3 \times 3}$  to be the block of  $V^{-1}$  defined by the  $3^2$  elements of Eq. 6.3 for  $i, i' \in \{1, 2, 3\}$ . Define  $V_\alpha = V_{\alpha,\alpha}^{-1} \in \mathbb{R}^{3 \times 3}$  which is a  $3 \times 3$  diagonal block of  $V^{-1}$ .

**Evaluation of  $\bar{\rho}_q(x)$  and  $R_{\rho_q, \rho_q}(x_a, x_b)$  from the pdf on the microscopic degrees of freedom** From Eq. 6.10 it follows that

$$\bar{\rho}_q(x) = \sum_{\alpha=1}^{N_p} f_\alpha \int_{\delta_\alpha \in \mathbb{R}^3} N(\mathbf{0}_3, q^2 I_3)(x - (x_{0,\alpha} + \delta_\alpha)) N(\mathbf{0}_3, \beta V_\alpha)(\delta_\alpha) d^3 \delta_\alpha \quad (6.15)$$

$$= \sum_{\alpha=1}^{N_p} f_\alpha \frac{1}{(2\pi)^{3/2}} \frac{1}{\sqrt{\det(\beta V_\alpha + q^2 I_3)}} \times \\ \times \exp\left(-\frac{1}{2q^4}(x - x_{0,\alpha})^T \left(q^2 I_3 - (q^{-2} I_3 + (\beta V_\alpha)^{-1})^{-1}\right)(x - x_{0,\alpha})\right) \quad (6.16)$$

where the general form of the integral of the product of two  $N_x$ -dimensional Gaussians which is used to derive Eq. 6.16 is

$$\int N(m_1, Q_1)(x) N(m_2, Q_2)(x) dx \\ = \frac{1}{(2\pi)^{N_x/2}} \frac{1}{\sqrt{\det(Q_2 + Q_1)}} \exp\left(-\frac{1}{2}\left(m_1^T Q_1^{-1} m_1 + m_2^T Q_2^{-1} m_2 \right. \right. \\ \left. \left. - (Q_1^{-1} m_1 + Q_2^{-1} m_2)^T (Q_1^{-1} + Q_2^{-1})^{-1} (Q_1^{-1} m_1 + Q_2^{-1} m_2)\right)\right). \quad (6.17)$$

From Eq. 6.12 it follows that

$$R_{\rho_q, \rho_q}(x_a, x_b) = \sum_{\alpha=1}^{N_p} \sum_{\alpha'=1}^{N_p} f_\alpha f_{\alpha'} \int_{\delta_\alpha \in \mathbb{R}^3} \int_{\delta_{\alpha'} \in \mathbb{R}^3} N(\mathbf{0}_3, q^2 I_3)(x_a - (x_{0,\alpha} + \delta_\alpha)) \times$$

$$\times N(\mathbf{0}_3, q^2 I_3)(x_b - (x_{0,\alpha'} + \delta_{\alpha'}))N(\mathbf{0}_{2.3}, \beta W_{\alpha,\alpha'}) \begin{pmatrix} \delta_\alpha \\ \delta_{\alpha'} \end{pmatrix} d^3 \delta_\alpha d^3 \delta_{\alpha'} \quad (6.18)$$

where

$$W_{\alpha,\alpha'} = \begin{bmatrix} V_\alpha & V_{\alpha,\alpha'} \\ V_{\alpha',\alpha} & V_{\alpha'} \end{bmatrix} \in \mathbf{R}^{2.3 \times 2.3}. \quad (6.19)$$

If  $V_{\alpha,\alpha'} = \mathbf{0}_{3,3}$  then  $V_{\alpha',\alpha} = V_{\alpha,\alpha'} = \mathbf{0}_{3,3}$  also. In this case,  $\delta_\alpha$  and  $\delta_{\alpha'}$  are independent, their joint pdf factors, and the value of  $R_{\rho_q, \rho_q}(x_a, x_b)$  is the product of two values of  $\bar{\rho}_q(\cdot)$ , i.e.,

$$R_{\rho_q, \rho_q}(x_a, x_b) = \bar{\rho}_q(x_a) \bar{\rho}_q(x_b). \quad (6.20)$$

If  $V_{\alpha,\alpha'} \neq \mathbf{0}_{3,3}$  then the integral is truly six dimensional and can be written in the form

$$\begin{aligned} R_{\rho_q, \rho_q}(x_a, x_b) &= \sum_{\alpha=1}^{N_p} \sum_{\alpha'=1}^{N_p} f_\alpha f_{\alpha'} \int_{\delta_\alpha \in \mathbf{R}^3} \int_{\delta_{\alpha'} \in \mathbf{R}^3} N \left( \begin{bmatrix} x_a - x_{0,\alpha} \\ x_b - x_{0,\alpha'} \end{bmatrix}, q^2 I_{2.3} \right) \begin{pmatrix} \delta_\alpha \\ \delta_{\alpha'} \end{pmatrix} \times \\ &\quad \times N(\mathbf{0}_{2.3}, \beta W_{\alpha,\alpha'}) \begin{pmatrix} \delta_\alpha \\ \delta_{\alpha'} \end{pmatrix} d^3 \delta_\alpha d^3 \delta_{\alpha'} \quad (6.21) \\ &= \sum_{\alpha=1}^{N_p} \sum_{\alpha'=1}^{N_p} f_\alpha f_{\alpha'} \frac{1}{(2\pi)^{3/2}} \frac{1}{\sqrt{\det(\beta W_{\alpha,\alpha'} + q^2 I_{2.3})}} \times \\ &\quad \times \exp \left( -\frac{1}{2q^4} \begin{bmatrix} x_a - x_{0,\alpha} \\ x_b - x_{0,\alpha'} \end{bmatrix}^T \left( q^2 I_{2.3} - (q^{-2} I_{2.3} + (\beta W_{\alpha,\alpha'})^{-1})^{-1} \right) \begin{bmatrix} x_a - x_{0,\alpha} \\ x_b - x_{0,\alpha'} \end{bmatrix} \right) \end{aligned} \quad (6.22)$$

where Eq. 6.17 was again used.

**Use of  $\bar{\rho}_q(x)$  and  $R_{\rho_q, \rho_q}(x_a, x_b)$  to evaluate the mean and covariance of the coefficients in an orthonormal expansion of the electron scattering intensity** In the formulas of this section it is assumed that there is only one discrete class of object. The generalization to multiple classes with the statistics conditional on the class requires only the addition of a class label to the notation. The object

is represented by an orthonormal expansion with real-valued basis functions which are denoted by  $\phi_\tau(x)$  indexed by  $\tau$  as described in Eqs. ??–?? where generalizations to complex-valued basis functions are straightforward. This is a fairly general situation since it includes voxel basis functions,

$$\phi_{n_1, n_2, n_3}(x) = \begin{cases} 1/\Delta^{3/2}, & n_1\Delta \leq x_1 < (n_1 + 1)\Delta, \\ & n_2\Delta \leq x_2 < (n_2 + 1)\Delta, \\ & n_3\Delta \leq x_3 < (n_3 + 1)\Delta \\ 0, & \text{otherwise} \end{cases} \quad (6.23)$$

where  $\Delta$  is the real-space sampling interval and  $x_1$ ,  $x_2$ , and  $x_3$  are the components of  $x$ ; spherical harmonic basis functions,

$$\phi_{l,n,p}(x) = \begin{cases} \Psi_{l,n}(\theta(x), \varphi(x))h_{l,p}(\|x\|), & \|x\| \leq R \\ 0, & \text{otherwise} \end{cases} \quad (6.24)$$

where  $\Psi_{l,n}$  is an appropriate linear combination of spherical harmonics of order  $l$  and  $h_{l,p}$  is a radial basis function, perhaps a linear combination of spherical Bessel functions of the first kind of order  $l$ , and  $\theta(x)$  and  $\varphi(x)$  are the spherical angles of  $x$ ; and so forth.

Let  $c$  be a vector with components  $c_\tau$ . The goal is to compute the mean, correlation, and covariance of  $c$ , which are denoted by  $m_c$ ,  $R_{c,c}$ , and  $C_{c,c}$ , respectively, and defined by  $m_c = E[c]$ ,  $R_{c,c} = E[cc^T]$ , and  $C_{c,c} = E[(c - m_c)(c - m_c)^T]$ , respectively. Take the expectation of Eq. ?? with respect to the microscopic degrees of freedom which are hidden in  $\rho(x)$  to find that the  $\tau$ th component of  $m_c$  is

$$(m_c)_\tau = E[c_\tau] = \int_{x \in \mathbb{R}^3} \bar{\rho}_q(x) \phi_\tau(x) d^3x. \quad (6.25)$$

Similarly, take the expectation of the product of Eq. ?? with itself at two possibly different values of the index ( $\tau$  and  $\tau'$ ) with respect to the microscopic degrees

of freedom which are hidden in  $\rho(x)$  to find that the  $(\tau, \tau')$ th element of  $R_{c,c}$  is

$$(R_{c,c})_{\tau,\tau'} = E[c_\tau c_{\tau'}] = \int_{x \in \mathbb{R}^3} \int_{x' \in \mathbb{R}^3} R_{\rho_q, \rho_q}(x, x') \phi_\tau(x) \phi_{\tau'}(x') d^3x d^3x'. \quad (6.26)$$

The details of the evaluation of these integrals depends on the specific choice of basis functions. Finally,

$$C_{c,c} = R_{c,c} - m_c m_c^T \quad (6.27)$$

which follows directly from the definition of  $C_{c,c}$ .

### 6.5.1 Estimates of the mean and the auto-correlation of the electron scattering intensity from the data

In the formulas of this section it is assumed that there is only one discrete class of object. The generalization to multiple classes with the statistics conditional on the class requires only the addition of a class label to the notation. The result of the models, algorithms, and software of Refs. [124, 125, 126] is estimates of the mean and the covariance of the vector  $c$  of coefficients that describes the object (Eq. ??). Let the estimates be denoted by  $\hat{m}_c$  and  $\hat{C}_{c,c}$ , respectively. Then a natural estimate of the correlation is  $\hat{R}_{c,c} = \hat{C}_{c,c} - \hat{m}_c \hat{m}_c^T$ . Suppose that the estimates are exact. Then it can easily be shown [125] that the mean, covariance, and auto-correlation of the electron scattering intensity, denoted by  $\hat{\rho}(x)$ ,  $\hat{C}_{\rho,\rho}(x_a, x_b)$ , and  $\hat{R}_{\rho,\rho}(x_a, x_b)$ , respectively, are

$$\hat{\rho}(x) = \sum_{\tau} (\hat{m}_c)_{\tau} \phi_{\tau}(x) \quad (6.28)$$

$$\hat{C}_{\rho,\rho}(x_a, x_b) = \sum_{\tau} \sum_{\tau'} (\hat{C}_{c,c})_{\tau,\tau'} \phi_{\tau}(x_a) \phi_{\tau'}(x_b) \quad (6.29)$$

$$\hat{R}_{\rho,\rho}(x_a, x_b) = \hat{C}_{\rho,\rho}(x_a, x_b) + \hat{\rho}(x_a) \hat{\rho}(x_b). \quad (6.30)$$

## 6.6 Estimating the mechanical properties from the data

The mechanical properties are the masses of the point masses ( $m_\alpha$  which define the matrix  $T$  of the kinetic energy quadratic form), the electron scattering intensities of the point masses ( $f_\alpha$ ), the equilibrium locations of the point masses ( $x_{0,\alpha}$ ), and the elements of the matrix  $V$  of the potential energy quadratic form where  $\alpha$  indexes the masses. In addition, there is the smoothing parameter  $q^2$  introduced in Eq. 6.8 and the inverse temperature  $\beta$  introduced in Section 6.4.

As noted in Section 6.5, the matrix  $T$  does not effect the statistical mechanical results because  $\bar{\rho}_q(x)$  and  $R_{\rho_q, \rho_q}(x_a, x_b)$  are functions of the microscopic variables  $\delta$  but not  $\dot{\delta}$ . A natural approximation is to assume that the mass  $m_\alpha$  and the electron scattering intensity  $f_\alpha$  of the  $\alpha$ th point mass are proportional. Let  $\zeta$  be the proportionality constant, i.e.,

$$m_\alpha = \zeta f_\alpha, \quad (6.31)$$

which will occur in all of the results of these calculations.

Since many electron micrographs are not calibrated, the electron scattering intensity can only be determined up to an arbitrary scaling or, equivalently, in unknown units.

Because  $\beta$  and  $V$  only occur in the combination  $\beta V$  and a typical experiment flash freezes the specimen from only a single initial temperature, it is not possible to separately determine  $\beta$  and  $V$ . In a biological experiment, this is a natural constraint since the object may not be stable at other than its natural temperature. Therefore,  $\beta$  will be set to 1 and  $V$  will be determined.

The smoothing parameter  $q^2$  is a function of the quality of the data and will

be determined from the spatial resolution achieved in the 3-D reconstruction of the object.

The remaining statistical mechanical parameters are the electron scattering intensities of the point masses ( $f_\alpha$ ), the equilibrium locations of the point masses ( $x_{0,\alpha}$ ), and the elements of the matrix  $V$  of the potential energy quadratic form where  $\alpha$  indexes the masses. One possible method for estimating these parameters is to select parameters to match the mean vector and covariance matrix of the coefficients  $c$  in the orthonormal expansion description of the object as computed from the statistical mechanical model versus from the data, i.e., match the mean vectors by matching Eq. 6.25 and  $\hat{m}_c$  (computed from the data) and match the covariance matrices by matching Eq. 6.27 (computed via Eqs. 6.25 and 6.26) and  $\hat{C}_{c,c}$  (computed from the data).

In order to apply this method, it is necessary to define measures of difference for the mean vector and the covariance matrix. For the mean vector, the most straightforward measure of difference is the Euclidean norm  $\|\cdot\|$ . Minimizing this measure of difference is solving a mixed linear and nonlinear least squares problem since  $\bar{\rho}_q(x)$  (Eq. 6.16) and therefore  $m_c$  (Eq. 6.25) is linear in the  $f_\alpha$  parameters. This is a desirable situation since there are special numerical algorithms for this type of problem [40, 57, 9]. For the covariance matrix, there are many matrix norms that could be applied to the difference of the two matrices. The Euclidean or Frobenius norm [68, Exercise 3, p. 358] is probably the most straightforward measure of difference since for any matrix  $A \in \mathbb{C}^{n \times n}$  with elements  $a_{i,j}$  the Euclidean norm is  $\|A\|_E = (\sum_{i=1}^n \sum_{j=1}^n |a_{i,j}|^2)^{1/2}$  so minimizing the Euclidean norm is a nonlinear least squares problem. (The parameters  $f_\alpha$  no longer enter linearly but instead enter quadratically). Alternatives are the ma-

trix norms induced by the Euclidean (i.e.,  $l_2$ ), the  $l_1$ , and the  $l_\infty$  vector norms [68, Exercises 5, 9, and 10; p. 365–366] but the computation of these norms is more difficult than that of the Euclidean matrix norm. The natural way to combine the two problems of matching the mean and matching the covariance is to add the two norms with a weight on one of the norms where the weight is chosen empirically.

An alternative way is to measure the difference between the two probability density functions (pdfs) implied by the mean vectors and covariance matrices. The mathematical model used in Refs. [124, 125, 126] uses a Gaussian pdf for the vector  $c$  of parameters. While  $\hat{m}_c$  and  $\hat{C}_{c,c}$  are only estimates of the parameters in the Gaussian pdf, it is still natural to take  $N(\hat{m}_c, \hat{C}_{c,c})(\cdot)$  as one of the two pdfs. The statistical mechanical model predicts a pdf on  $c$ , but it is not Gaussian. However, since all that is computed is the mean and covariance of the pdf, it is still natural to take  $N(m_c, C_{c,c})(\cdot)$  as the second pdf. A natural way in which to compare two pdfs is the Kullback-Leibler divergence [66][22, Eq. 2.27, p. 18] which, if the two pdfs are denoted by  $p$  and  $q$ , is denoted by  $D_{\text{KL}}(p||q)$  and defined by

$$D_{\text{KL}}(p||q) = \int p(x) \ln \left( \frac{p(x)}{q(x)} \right) dx. \quad (6.32)$$

The divergence is not symmetrical, i.e.,  $D_{\text{KL}}(p||q) \neq D_{\text{KL}}(q||p)$ , and  $p$  of  $D_{\text{KL}}(p||q)$  is typically the “true” pdf while  $q$  is approximating  $p$ . Therefore,  $p(c) = N(\hat{m}_c, \hat{C}_{c,c})(c)$  and  $q(c) = N(m_c, C_{c,c})(c)$ . Minimization of Eq. 6.32 for this case with respect to the statistical mechanical parameters determines all of the parameters jointly while matching, in a pdf-sense, the mean vectors and covariance matrices. While Eq. 6.32 appears to be an integral over a number of dimensions equal to the number of coefficients in the orthonormal expansion, which is large, for the case of two Gaussian pdfs, Eq. 6.32 can be computed symboli-



cally: if  $p(x) = N(m_p, Q_p)(x)$  and  $q(x) = N(m_q, Q_q)(x)$  then

$$D_{\text{KL}}(p||q) = \frac{1}{2} \left[ \ln \det(Q_p^{-1} Q_q) + \text{tr} [Q_p Q_q^{-1}] - n + (m_p - m_q)^T Q_q^{-1} (m_p - m_q) \right] \quad (6.33)$$

where  $n$  is the dimension of the random vectors described by the pdfs.

## 6.7 A model for $\mathcal{V}(\delta)$ and $V$

Perhaps something is known about the potential energy function  $\mathcal{V}(\delta)$ . For instance, at atomic resolution the potential energy function from a molecular dynamics system could be used to define which elements in the matrix  $V$  of Eq. 6.3 are nonzero. At a lower spatial scale, perhaps the biologist user has ideas about how the components of the object move which could lead to definitions of point masses and nonzero elements in the matrix  $V$  of Eq. 6.3 that are specific to that particular biological system.

A third way in which to structure  $\mathcal{V}$  and  $V$  is to think of the point mass model as a discretization of a continuum model. In particular, suppose that the equilibrium has a regular cubic array of point masses so that

$$\alpha = (\alpha_1, \alpha_2, \alpha_3)^T \quad (6.34)$$

$$x_{0,\alpha} = (\alpha_1 \Delta, \alpha_2 \Delta, \alpha_3 \Delta)^T \quad (6.35)$$

where  $\Delta \in \mathbb{R}$  is the sampling interval and  $\alpha_1, \alpha_2, \alpha_3 \in \mathcal{Z}$ . Suppose further that each point mass interacts with only its 6 nearest neighbors, that the interactions are via linear springs, and that the interactions are not frustrated, i.e., the ground state of the entire system is the ground state of each interaction. Then the general linear spring potential energy function of Eq. 6.5 can be simplified:

The set of indices  $\mathcal{N}_\alpha$  contains the 6 index vectors

$$\mathcal{N}_\alpha = \left\{ \begin{bmatrix} \alpha_1 \pm 1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}, \begin{bmatrix} \alpha_1 \\ \alpha_2 \pm 1 \\ \alpha_3 \end{bmatrix}, \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \pm 1 \end{bmatrix} \right\}, \quad (6.36)$$

the sum  $\sum_{\alpha' \in \mathcal{N}_\alpha}$  has 6 terms, and the rest lengths  $b_{\alpha, \alpha'}$  are all equal to the sampling interval, i.e.,

$$b_{\alpha, \alpha'} = \Delta \quad (6.37)$$

with the result that

$$\mathcal{V}_{\mathbf{x}_0}(\delta) = \sum_{\alpha_1=1}^{\sqrt[3]{N_p}} \sum_{\alpha_2=1}^{\sqrt[3]{N_p}} \sum_{\alpha_3=1}^{\sqrt[3]{N_p}} \sum_{\alpha' \in \mathcal{N}_\alpha} \frac{1}{2} k_{\alpha, \alpha'} \left[ \left\| \begin{bmatrix} \alpha_1 \Delta \\ \alpha_2 \Delta \\ \alpha_3 \Delta \end{bmatrix} + \delta_\alpha \right\| - \left\| \begin{bmatrix} \alpha'_1 \Delta \\ \alpha'_2 \Delta \\ \alpha'_3 \Delta \end{bmatrix} + \delta_{\alpha'} \right\| - \Delta \right]^2 \quad (6.38)$$

where  $\alpha = (\alpha_1, \alpha_2, \alpha_3)^T$  and  $\alpha' = (\alpha'_1, \alpha'_2, \alpha'_3)^T$ . This is attractive because the equilibrium positions  $x_{0, \alpha'}$  which enter the predictions of the statistical mechanics in nonlinear ways, are now known and therefore are not computed from the data. In particular, only the  $N_p$  electron scattering intensities  $f_\alpha$  and the  $6N_p$  spring constants  $k_{\alpha, \alpha'}$  must be computed from the data. It is likely helpful to restrict how rapidly the spring constants  $k_{\alpha, \alpha'}$  change as  $\alpha$  moves across the lattice so that the changes are appropriate for the spatial resolution of the reconstruction. This can be done by a variety of regularization methods. It is likely that an iterative algorithm can be derived in which each iteration involves first an update of the  $f_\alpha$  parameters and second an update of the  $k_{\alpha, \alpha'}$  parameters using the new  $f_\alpha$  parameters. Such an algorithm might be attractive because the  $f_\alpha$  parameters occur linearly in the predictions of the means  $\bar{\rho}_q(x)$  (Eq. 6.16) and  $m_c$  (Eqs. 6.25 and 6.16) so the update of the  $f_\alpha$  parameters, if it is based solely on  $\bar{\rho}_q(x)$  or solely on  $\bar{\rho}_q(x)$ , is straightforward.

## 6.8 Characteristic frequencies and normal modes

With a complete set of statistical mechanical parameters, the characteristic frequencies and normal modes can be computed.

**Standard theory based on known  $T$  and  $V$  [68, Section 5.12]** Let the second time derivative of  $\delta_\alpha$  be denoted by  $\ddot{\delta}_\alpha$ . Let  $\ddot{\delta} = (\ddot{\delta}_1^T, \dots, \ddot{\delta}_{N_p}^T)^T$ . With the quadratic definitions of kinetic (Eq. 6.1) and potential (Eq. 6.4) energy, the dynamical equations can be derived by many methods with the conclusion that

$$T\ddot{\delta} + V\delta = \mathbf{0}_{3N_p}. \quad (6.39)$$

It is anticipated that the solution is of the form

$$\delta(t) = \mathbf{u} \exp(i\omega t) \quad (6.40)$$

where  $\omega \in \mathbf{R}$  is a characteristic frequency,  $t \in \mathbf{R}$  is time, and  $i = \sqrt{-1}$ . Substituting Eq. 6.40 into Eq. 6.39 implies that

$$(-\lambda T + V)\mathbf{u} = \mathbf{0}_{3N_p} \quad (6.41)$$

where  $\lambda = \omega^2$  which is a generalized eigen problem where  $\lambda$  is the value and  $\mathbf{u}$  is the vector. If  $H$  is a Hermitian positive semidefinite matrix then it has a unique Hermitian positive semidefinite square root, denoted by  $H^{1/2}$ , such that  $H^{1/2}H^{1/2} = H$  and similarly for semidefinite replaced by definite [68, Proposition 1, pp. 181–182]. For any Hermitian matrix  $H$ ,  $H = UDU^*$  where  $D$  is diagonal and real and  $U$  is unitary, which is the eigen vector eigen value decomposition of  $H$ . Then, for semidefinite Hermitian  $H$ ,  $H^{1/2} = UD_{1/2}U^*$  where  $D_{1/2}$  is the diagonal and real matrix which has elements that are the square roots of the elements of  $D$ . Apply these results to  $T$ , which is positive definite, to find

that there is a square root of  $T$ , denoted by  $T^{1/2}$ , which is also positive definite and therefore has an inverse, denoted by  $T^{-1/2}$ . Multiply Eq. 6.41 on the left and right by  $T^{-1/2}$  to get

$$(-\lambda I_{3N_p} + T^{-1/2} V T^{-1/2}) \mathbf{u} = \mathbf{0}_{3N_p} \quad (6.42)$$

which is an eigen problem for  $T^{-1/2} V T^{-1/2}$ . Since  $T^{-1/2} V T^{-1/2}$  is Hermitian positive semidefinite (definite if  $V$  is definite), the answer is that the eigen values are real and non negative (group them into a diagonal matrix  $\Lambda$ ), the eigen vectors can be taken as orthonormal (group them into a corresponding unitary matrix  $U$ ), and  $T^{-1/2} V T^{-1/2} U = U \Lambda$  or equivalently

$$U^* T^{-1/2} V T^{-1/2} U = \Lambda \quad (6.43)$$

since  $U$  is unitary. Because  $V$  and  $T$  are real,  $U$  is actually a real orthogonal matrix rather than a potentially complex unitary matrix.

The characteristic frequencies are the square roots of the eigenvalues  $\lambda$ , i.e.,

$$\omega_j = \sqrt{\lambda_j}. \quad (6.44)$$

The normal modes are defined by the linear transformation from  $\delta(t)$  to  $\xi(t)$  defined by

$$\delta(t) = T^{-1/2} U \xi(t) \quad (6.45)$$

which has the following properties. First, the dynamical equations decouple: Substitute Eq. 6.45 into Eq. 6.39 to get  $T T^{-1/2} U \ddot{\xi} + V T^{-1/2} U \xi = \mathbf{0}_{3N_p}$  which implies  $T^{1/2} U \ddot{\xi} + V T^{-1/2} U \xi = \mathbf{0}_{3N_p}$ . Multiply through on the left by  $U^* T^{-1/2}$  to get, using  $U^* U = I_{3N_p}$ ,  $\ddot{\xi} + U^* T^{-1/2} V T^{-1/2} U \xi = \mathbf{0}_{3N_p}$  and apply Eq. 6.43 to get

$$\ddot{\xi} + \Lambda \xi = \mathbf{0}_{3N_p}. \quad (6.46)$$

Because  $\Lambda$  is a diagonal matrix, the different components of  $\xi$  do not interact. Second, and equivalently, the two energy quadratic forms (Eqs. 6.1 and 6.4) are

diagonal:

$$\mathcal{T}(\dot{\delta}) = \frac{1}{2} \dot{\delta}^T T \dot{\delta} = \frac{1}{2} (T^{-1/2} U \dot{\xi})^T T T^{-1/2} U \dot{\xi} = \frac{1}{2} \dot{\xi}^T U^* T^{-1/2} (T^{1/2} T^{1/2}) T^{-1/2} U \dot{\xi} = \frac{1}{2} \dot{\xi}^T \dot{\xi} \quad (6.47)$$

and

$$\mathcal{V}_{x_0}(\delta) = \frac{1}{2} \delta^T V \delta = \frac{1}{2} (T^{-1/2} U \xi)^T V T^{-1/2} U \xi = \frac{1}{2} \xi^T U^* T^{-1/2} V T^{-1/2} U \xi = \frac{1}{2} \xi^T \Lambda \xi \quad (6.48)$$

by Eq. 6.43.

By the non-interacting characteristic shown in Eq. 6.46, if only the  $j$ th component of  $\xi$  is non zero at the initial condition, then only the  $j$ th component is ever non zero. In terms of the original variables  $\delta$ , only the  $j$  component of  $\xi$  is non zero in the initial condition if and only if the initial condition for  $\delta$  is proportional to the  $j$ th column of  $T^{-1/2} U$  (Eq. 6.45). This is called the  $j$ th normal mode and, if  $\mathbf{w}_j$  is the  $j$ th column of  $T^{-1/2} U$ , then the time evolution of the  $j$ th normal mode is (Eq. 6.40)

$$\delta_j(t) = \mathbf{w}_j \exp(i\omega_j t). \quad (6.49)$$

## 6.9 Practical Computation

### 6.9.1 Equilibrium Positions Determination from Discretizing the continuous 3D Intensity

To build a mechanical model with point masses on a dense regular lattice implies that the number of springs connecting the point masses will be huge, (i.e. for image on the order of  $10^2$ , we will need roughly  $2.7 \times 10^7$  spring constants),

thus making parameters estimation of the model infeasible. It is probably not necessary to build a model based on a cartesian grid of point masses, especially when the goal is a model that captures the overall shape of a macromolecular structure, which is essential in predicting structural dynamics. Thus, it is beneficial to select equilibrium positions of point masses based on the shape and density value at each location of a specific structure. This becomes a problem of how to discretize a continuous 3-D density map into  $N$  discrete points that best represent the shape of the structure. Techniques in data compressions such as vector quantization can be used to solve such a problem, in which the 3-D density map is separate into  $N$  voronoi cells and the centroids of these cells are used as the equilibrium positions of point masses for the mechanical model. These points have the nice property that the distance between any points within a cell and its centroid is smallest, thus making these points representing the overall shape of the 3D density well. [6] [83] . Using the vector quantization technique, typically the overall shape of macromolecular structure can be represented pretty well by  $10^2$  to  $10^4$  discrete points. Since the equilibrium positions of point masses are not on a regular grid, a criterion is needed to specify how point masses are connected in the model. A natural way to connect "neighbor" point masses is by measuring their pairwise distance. We define a specific threshold on the distance such that we connect each pair of point masses when their distance is within the threshold. This parameter is crucial for the model and might be different when  $N$  is different. In particular, if it is too small, then the structure is not well connected and one is not able to predict any coherent motion from it. On the other hand, if it is too large, any structure will be restrained to exhibit some form coherent motion that might not reflect the actual biological meaningful motion of the structure [59, 60]. Thus, it will require a validation procedure

to determine experimentally a range of this distance threshold that works well. Among the literature on network models, it is suggested that the distance is set between 8Å to 16Å in order for the model to provide a reasonable result [92, 108].

### 6.9.2 Strategies for Computing the Eigensystem of $V$

As discussed in Section 6.6, the model parameters can be estimated by finding the set of parameters that match the predict mean and covariance intensity from our model with the experimental ones estimated from cryo-EM images. Since the spring constants are the key model parameters that influence the structural dynamics, we will focus on the estimation of the spring constants from image statistics [29]. Let  $\mathbf{k}$  be the set of spring constants for our model,  $\bar{\rho}$  and  $\bar{\Sigma}$  be the experimental mean and covariance from experimental images and  $\hat{\rho}$  and  $\hat{\Sigma}$  be the predictions from our model. Then, the best estimate of  $\mathbf{k}$  is the solution to the following constraint optimization problem:

$$\begin{aligned} \min_{\mathbf{k}} \quad & \|\hat{\rho}(\mathbf{k}) - \bar{\rho}\|_2 + \lambda \|\hat{\Sigma}(\mathbf{k}) - \bar{\Sigma}\|_F \\ \text{s.t.} \quad & k_i \geq 0 \end{aligned} \tag{6.50}$$

in which  $\lambda$  is a weight that determine how important the difference of mean and covariance in this estimation is.

To solve the constraint minimization problem, the active-set algorithm in Matlab is used, which requires the error function value and its gradient at the current value of  $\mathbf{k}$  in each step. No analytical form of the gradient of the cost function is available, so a finite difference approximate, which requires repeat evaluations of the expected mean and covariance using Eq. 6.10 and Eq. 6.22

at each step, is computed. A key quantity in the computation is the inverse of the hessian matrix of the potential energy of a mechanical system. The hessian matrix,  $V$ , is rank deficient because the internal energy is invariant with respect to translations, resulting in  $V$  having a nontrivial null space. We computed a pseudo-inverse of  $V$  based on its Singular Value Decomposition (SVD). Since  $V$  is hermitian, the SVD is equivalent to its Eigenvalue Decomposition. Let  $\lambda_i$  be the non-zero eigenvalues of  $V$  in ascending order and  $v_i$  its corresponding eigenvectors. Then  $V^\dagger$ , the pseudo-inverse of  $V$ , is :

$$V^\dagger = \sum_{i=1}^{3N_p-9} \frac{1}{\lambda_i} v_i v_i^T \quad (6.51)$$

For a typical problem,  $N_p$  is on the order between  $10^2$  to  $10^4$  and the computation of eigenvalues and eigenvectors typically require  $O(N_p^4)$ , thus making this step the bottleneck in the overall calculation.

Two strategies are employed to reduce the computational load at this step. For median size problems (i.e.  $N_p \sim 10^2$ ), a fast method of computing the eigensystem of a rank-1 update to an already known eigensystem can be used. During each step of the minimization process, the gradient of the error function with respect to spring constant is computed via finite difference. In the calculation of the numerical gradient, a single spring constant is modified in succession. It can be shown that the hessian matrix with a specific spring constant taking on two different values, which we denoted  $V_1$  and  $V_2$ , respectively, can be related by a rank-1 update:

$$V_2 = V_1 + \alpha u u^T \quad (6.52)$$

where  $\alpha$ , a scalar, and  $u$ , a vector, depend on the difference between the two spring constant values. Suppose we have already computed the eigenvalues and eigenvectors of  $V_1$ , then there is an easy way to update the eigenvalues and



eigenvectors for  $V_2$  based on that of  $V_1$ . Let  $Q$  be an orthonormal matrix where the columns are eigenvectors of  $V_1$  and  $D$  be a diagonal matrix such that the diagonal elements are the eigenvalues of  $V_1$ , then

$$V_2 = V_1 + \alpha uu^T \quad (6.53)$$

$$= Q^T(D + \alpha zz^T)Q \quad (6.54)$$

$$= Q^T Q_2^T \text{diag}(\tilde{\lambda}_1, \dots, \tilde{\lambda}_n) Q_2 Q \quad (6.55)$$

where  $z = Qu$ , and  $Q_2^T \text{diag}(\tilde{\lambda}_1, \dots, \tilde{\lambda}_n) Q_2$  is the eigen-decomposition of  $D + zz^T$ . As proven in Ref. [41] Ch. 8, the eigenvalues of  $D + \alpha zz^T$  are the roots to the following simple polynomial,

$$f(\lambda) = 1 + \alpha \left( \frac{z_1^2}{d_1 - \lambda} + \dots + \frac{z_n^2}{d_n - \lambda} \right) \quad (6.56)$$

and the eigenvectors are simply multiples of  $(D - \lambda_i I)^{-1} z$ . There are numerical stability issues preventing the use of standard root-finding techniques when the roots of Eq. 6.56 are close to each other, thus, algorithms described in [16, 5] are implemented to obtain these roots accurately.

With this rank-1 update shortcut, we only need to compute the eigensystem using a standard method once for each step during the minimization process. Subsequent ones can be computed more efficiently using the method described above to obtain a numerical gradient of the error function.

For a large system space (i.e  $N_p > 10^3$ ), even with the rank-1 update procedure, the computation can quickly become infeasible, thus we seek only an approximation to the pseudo-inverse. Since  $\lambda_i$  in Eq. 6.51 is sorted in ascending order, the early terms contribute mostly significantly to the pseudo-inverse. Thus, a natural approximation of  $V^\dagger$  is to sum only the first  $n$  terms,

$$V^\dagger \approx \sum_{i=1}^n \frac{1}{\lambda_i} v_i v_i^T, \quad (6.57)$$

where  $n$  is determined so that  $\frac{\lambda_1}{\lambda_n}$  is smaller than some threshold  $\tau$ . From experience,  $n$  is typically much smaller than  $3N_p$  for a  $\tau$  of  $10^{-4}$ , especially when  $N_p$  is large. In this case, iterative methods such as the Lanczos algorithm can be used to obtain the smallest  $n$  eigenvalues and eigenvectors of  $V$  efficiently [41].

## 6.10 Feasibility of Model Parameter Estimation

To explore the feasibility of obtaining mechanical parameters from the covariance of cryo-EM image data, a simple 2-D mechanical model is constructed. The simple 2-D model is motivated by the biomolecule immunoglobulin IgG, which consists of two subunits called Fab arms that are joined together at a hinge site to a stem subunit known as Fc. It has been shown that the two Fab arms of the molecule exhibit great mobility, in which they flex back and forth around the center line along the stem part of the IgG molecule [11]. Thus, this is a good example to illustrate how the variability of cryo-EM data originates from significant conformational dynamics of a macromolecule.

### 6.10.1 The IgG model

The simple 2-D model consists of 64 point mass situated on a 20 by 20 rectangular grid. Each subunit is represented by 20 point masses. The mass of the point particles are assumed to be constant. Each particle is connected to its eight nearest neighbors. The values of the spring constant within each subunit are set to be a relatively stiff value of 5, reflecting the lack of degrees of freedom within each subunit. On the other hand, the spring constants at the hinge sites are set to be

much smaller than 1, reflecting the relative ease of movement between Fab arms and the Fc stems. The smoothing parameter in the electron scattering model is set to 0.3, so the electron scattering intensity of one point particle does not overlap much with the neighboring particles. The 2-D model is shown in Figure 6.1, in which the point mass are represented by spheres with radius proportional to their smoothing parameter and the springs are represent by lines connecting particles where the line thickness is proportional to the spring constant values.

Several instances of point mass positions for the 2-D IgG model are plotted in Figure 6.2. To illustrate how the point particles move relative to their equilibrium positions, the equilibrium positions (marked as open circles) are plotted together with the instantaneous positions (marked as filled circles). The instantaneous positions reflect the random nature of the fluctuation of each particle around its equilibrium position microscopically but, as a whole, the two Fab subunits show coherent movement that reflects domain motions. Basically, two general type of global movements can be seen from different realizations of the model: (1) the two arms are moving in opposite directions, flexing back and forth around the center line (Fig. 6.2 (a)) and (2) the two arms are move together in one direction (Fig. 6.2 (b)), which capture the essential motions of the IgG molecule rather well.

The theoretical mean and variance of the pixel values of image data calculated according to Eq. 6.16 and Eq. 6.22 are shown in Fig. 6.3. In order to see how the local mechanical properties at the hinge site of the model affect the pixel values globally, two different IgG models (Model 1 and Model 2) are constructed, in which the spring constant at the hinge sites is set to be 1 and 0.01 respectively. Comparing Fig. 6.3 (a) & (b), the mean pixel values for Model 1 are higher than

those of Model 2. This is expected, since Model 2 has a smaller spring constant at the hinge site than that of Model 1, so the two arms of Model 2 will move further and more frequently away from its equilibrium position than that of Model 1. As a result, the pixel variances of Model 2 are larger than Model 1 (Fig. 6.3 (c) & (d) ). Thus, the mean pixel values, which are generally regarded as the image data from cryo-EM data, do not tell the whole story of the biomolecules that are being imaged.

### 6.10.2 Model parameter estimation

In the current calculation, the weight,  $\lambda$ , in Eq. 6.50 is set to 1. In order to avoid trapping in local minima during the optimization, multiple initializations are used. Since the minimum value is known (i.e., 0), the algorithm only restarts when the cost function from previous step is greater than some prescribed threshold. A maximum of 300 initializations are used during the minimization process. Furthermore, there are eight spring constants for each point mass, so there are a large number of parameters that need to be estimated. In order to reduce the computational load and make the cost function smoother, a multi-step procedure similar to that used in Section 2.4.2 is used. Specifically, two steps are used in the current calculation. In the first step, spring constants connected to the particles are assumed to be the same, which reduces the number of parameters needed to be estimated by three-fold. To avoid trapping in the local minima, the multi-start algorithm described above is used until the cost function is below a certain threshold. Then in the second step, the answer obtained from the first step is used as the initialization, in which all spring constants are estimated together without enforcing the local consistency criterion. As shown

in Fig. 6.5, the estimations are quite accurate.

### 6.10.3 Inferring structural dynamics: NMA analysis

Normal mode analysis (NMA) is performed on the IgG Model 2. Examining the characteristic frequencies, the two slowest modes have characteristic frequency of 0 corresponding the horizontal and vertical translation of the whole system. The characteristic frequencies of Mode 3 to 6 have comparable magnitudes, and then there is a sharp increase to the frequency of Mode 7, which is 10 times that of Mode 3. It is known that only the slow modes contribute significantly to the collective motion of macromolecules [92], thus only the slowest four modes are shown in Figure 6.4, in which the displacement vector of each point mass in each mode is shown as a vector with its base at the equilibrium position of the point mass. Modes 3 and 4 correspond to the two Fab subunits moving in opposite directions from each other horizontally and vertically, while the stem part has no net movement in this mode. Mode 5 and 6 correspond to the two Fab subunits moving in the same directions. The major movements of the IgG model as seen in Fig. 6.2 can be characterized by a linear combination of these four modes. For example, the movement seen in Fig. 6.2 (a), in which the two arms flex outward and downward can be represented as a combination of Mode 3 and Mode 5, while the motion seen in Fig. 6.2(b), in which the two arms move both to one side, can be seen as a combination of Mode 4 and Mode 6. Therefore, NMA analysis is able to capture the essential motion of a macromolecular assembly quite accurately. Furthermore, point masses in each domain move collectively in each of the four slowest modes, suggesting that NMA analysis is extremely useful in identifying subunits that function together within a com-

plex macromolecular assembly, which might not be identified easily by simple visual inspection.

## **6.11 Validation of Structural Dynamic Prediction Based on NMA Analysis**

To illustrate the application of the mechanical model, a mass-spring model of the open form of adenylate kinase [107] is constructed. Adenylate kinase is an important biological enzyme that is known to exhibit hinge-like motion at its two arms to bind its substrate and the crystal structures of both its open and close forms are known, which provide the ground truth to compare with its dynamical information.

Since the goal is to test the applicability of a coarse-grained model in structural dynamic prediction, we did not build the mechanical model based on the atomic coordinates. Instead, electron scattering intensity is generated at a resolution of 10Å using the open form of the x-ray crystallographic structure. Then two hundred equilibrium positions are obtained by discretizing the continuous electron-scattering intensity using a vector quantization technique [107]. The mass-spring model is shown in Fig. 6.6. Since there is no experimental cryo-EM data from which to estimate experimental mean and covariance, the model parameters of the mechanical model are estimated by comparing predicted B-factors with experimental ones obtained from x-ray crystallography, in a similar approach as is discussed in Section 6.6. This is possible because the mechanical model is generative and many quantities, including the B-factors, can be computed.

The predicted mean and covariance of the electron scattering intensity is computed and the mean is plotted in Fig. 6.7, in which the surface is colored according to the variance information at that location. The variance information agrees with our understanding of adenylate kinase, in which high variance appears at its two arms, where large domain motions occur to facilitate substrate docking. On the other hand, the groove between the two arms and the regions below have relative low intensity variance, reflecting the hinge site of the structure.

Normal mode analysis is performed on the mechanical model. The lowest non-zero frequency mode is plotted in Fig. 6.8(a) along with the displacement vectors between the open and close form of the x-ray structures (Fig. 6.8(b)). The similarity between the two vector fields supports the claim that a coarsed-grain model is able to capture global motions of macromolecular complexes accurately once the model parameters are identified.

## **6.12 A Real 3D Example: Modeling the Flock House Virus Capsid**

Adenylate kinase is a small protein comparing to super-molecular complexes such as a virus particle. To test its applicability in a real situation, we try to construct a model of the Flock House Virus (FHV), for which projection images from cryo-EM data are available. 528 projection images of FHV from single particle cryo-EM imaging experiments are obtained from The Scripps Research Institute. The mean and covariance intensity information based on these images are estimated in a previous study (Fig. 6.9 (a)) [127]. Since FHV possesses icosah-

hedral symmetry, to simplify calculations, all model parameters are assumed to retain such symmetry. One asymmetric unit is extracted from the mean intensity reconstruction of the full FHV capsid. Then, we discretize the mean intensity of one asymmetric unit of FHV using 50 points by applying the vector quantization technique. The whole FHV capsid model is constructed by applying the 60 icosahedral symmetry rotation operators to the coordinates of the first 50 point masses (Fig. 6.9 (b)). Minimization procedures described in Section 6.9.2 are used to estimate the spring constants of the mechanical model. Unfortunately, the algorithm does not proceed to a reasonable minimum even after many iterations. The experimental covariance matrix and the predicted matrix based on the final step of the minimization process are plotted in Fig. 6.10 (a) and (b) respectively. Comparing the two covariance matrix, it can be seen that the predicted one has much stronger correlation among point masses than the actual experimental one.

The set of spring constants for the network model actually act indirectly on the predicted covariance through the hessian matrix of the potential energy of the system, whose inverse is required in the calculation of both the mean vector and the covariance matrix. The spring constant values and the spring connectivity both influence the hessian matrix  $V$ . In previous calculations, we focused on the spring constants and determined the connectivity by connecting neighboring point masses that are within a certain distance of each other. This probably puts too much constraint on the hessian matrix, and hence on the predicted covariance matrix, making it difficult for the optimization algorithm to find a set of spring constants that match the predicted covariance closely to the experimental covariance. To confirm this hypothesis, the covariance matrices from models with random spring constants are computed, of which two such instances are



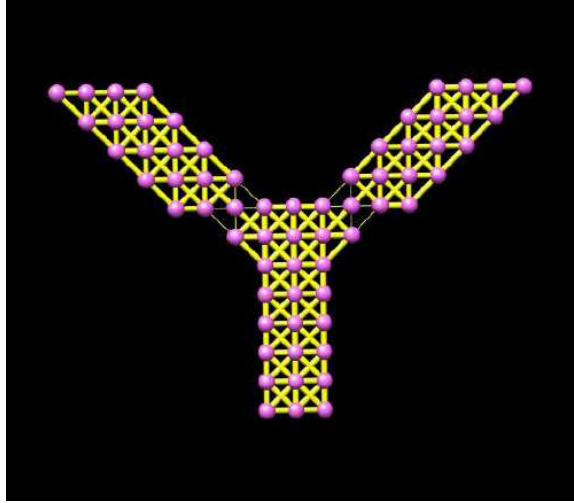


Figure 6.1: IgG model represented as spheres and rods. Each pink sphere represent one point mass of the model, connected by springs, represent as yellow rods with width proportional to the spring constant values. The spring constants are uniform everywhere in the IgG model except at the joints between the stem and the two arms which has a smaller spring constant value to allow the two arms to flex.

plotted in Fig. ??(a) and (b). Comparing the two covariance, they have surprisingly similar structures, even though the actual values at a specific element might be different. This hints at the fact that spring connectivity actually plays a strong role in determining the overall structure of the covariance matrix while the spring constants contribute in modifying specific values. Thus, it might be more appropriate to estimate both the spring constants and spring connectivity based on experimental data. However, it is hard to quantify the connectivity of a spring, thus, it might be easier to estimate the full hessian matrix  $V$ , which directly influences the mean vector and covariance matrix in a minimization procedure and then determine the spring constant and connectivity of the model based on this estimated hessian matrix.

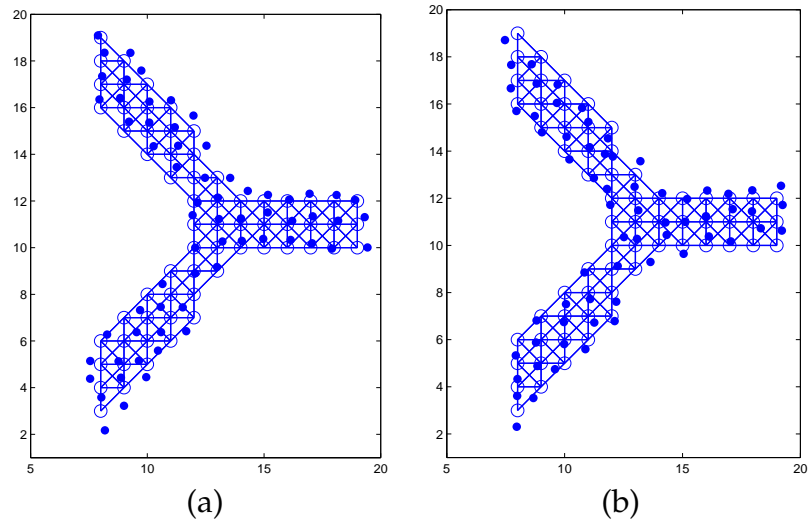


Figure 6.2: Two instances of the model. Open circles are the equilibrium positions and the filled circles are the instantaneous positions

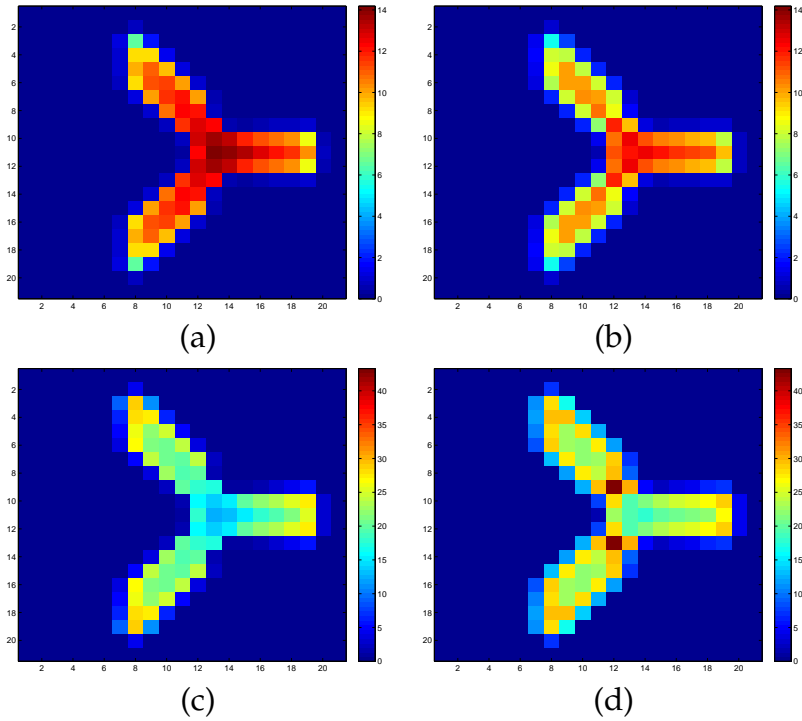


Figure 6.3: Mean and variance of the image pixel values of the IgG model

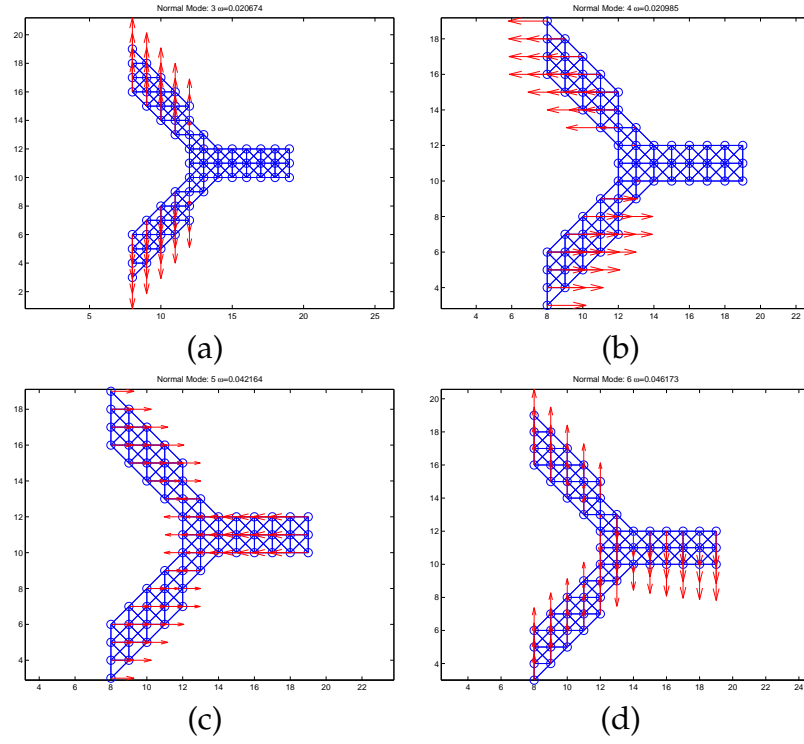


Figure 6.4: Mean and variance of the image pixel values of the IgG model

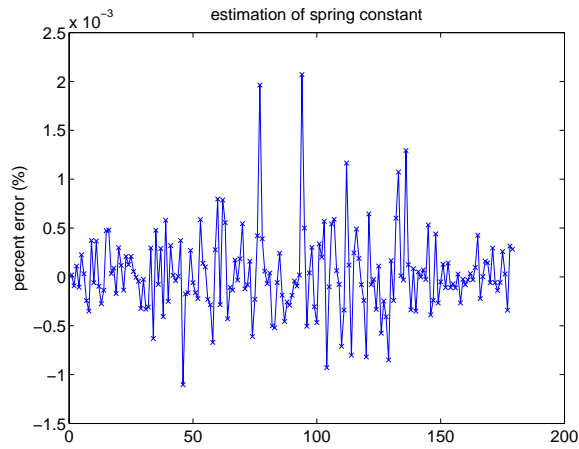


Figure 6.5: Percentage error in the estimation of the model parameters (spring constant) of the IgG model.

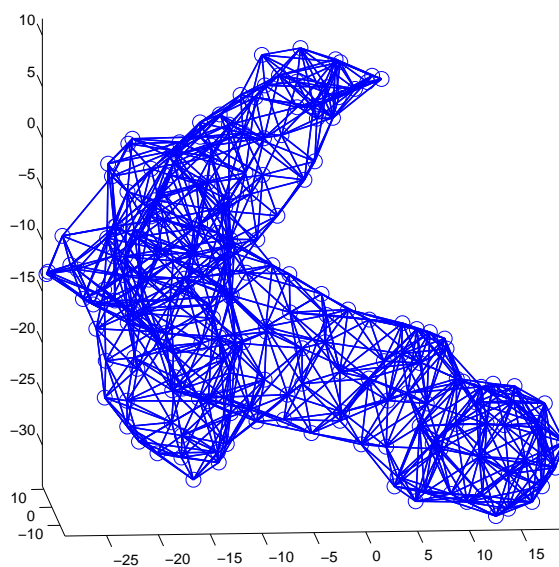


Figure 6.6: A plot of the mass-spring model of adenylate kinase. The equilibrium positions, which are represented by circles, are obtained from discretizing the density map of adenylate kinase which is generated synthetically from its x-ray crystallographic structure. A line connecting two circles represents a spring connecting two point masses.

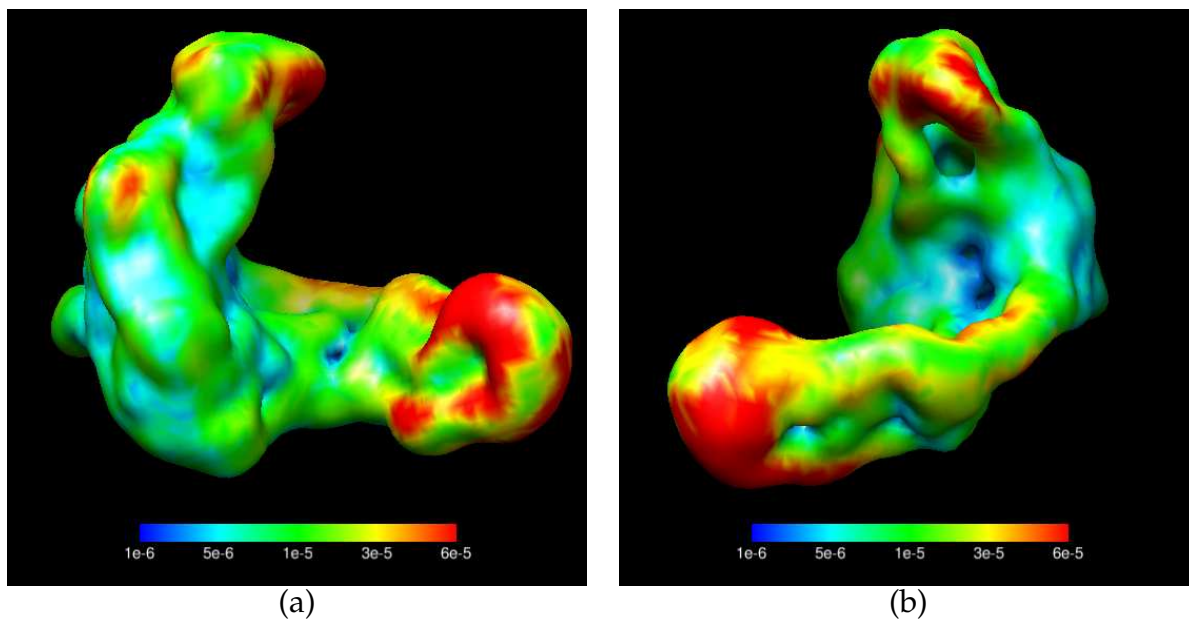


Figure 6.7: Variance of electron scattering intensity of adenylylate kinase as predicted from the model. The surface of the adenylylate structure is colored according to the variance information at that location. Red represents highest variance while blue represents the least variance Panel (a): A side view of the surface of adenylylate kinase, same orientation as the spring-mass model plot. Panel (b): A top-down view of the surface of the adenylylate kinase structure

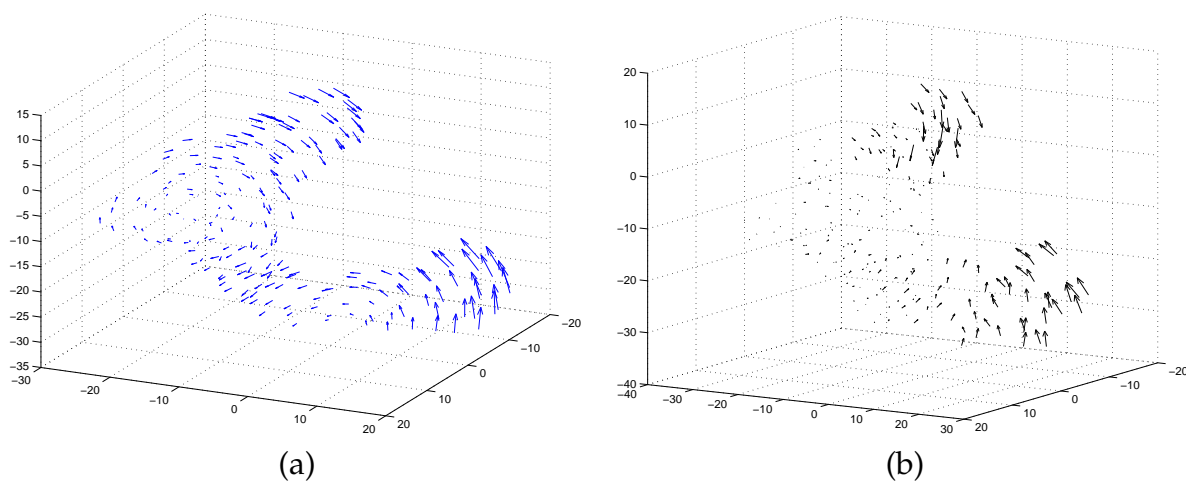
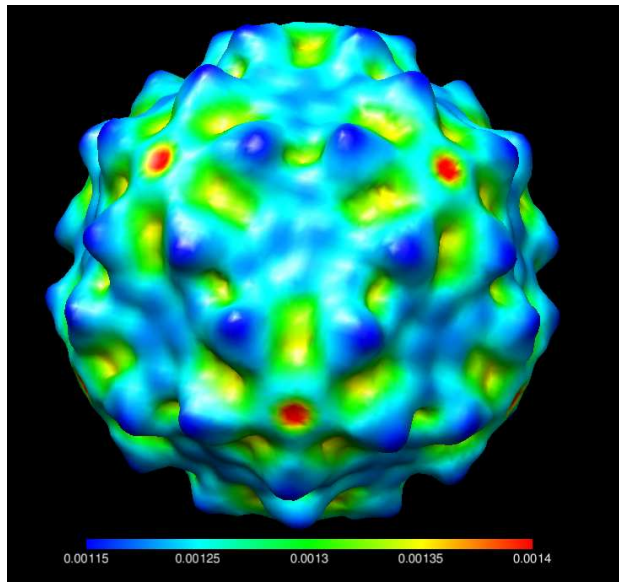
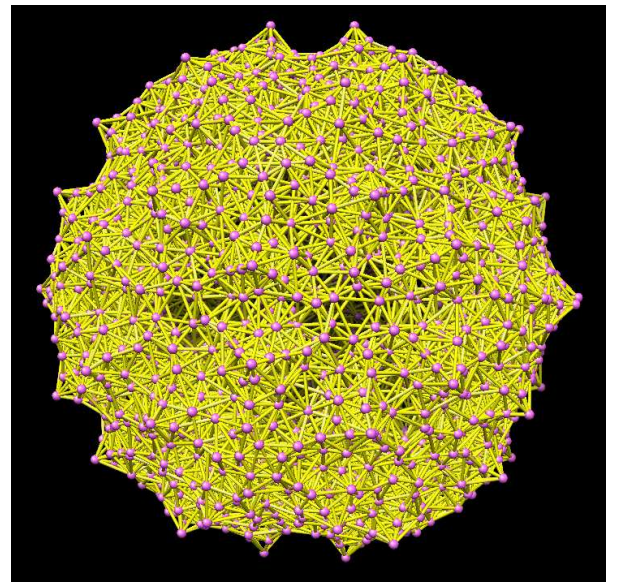


Figure 6.8: The open/close forms of adenylate kinase. Panel (a): The first normal mode of the 3-D mass-spring model from the 3-D adenylate kinase density map. Panel (b): Displacement vectors between the open and close form of the x-ray crystallographic structure of adenylate kinase. The similarity of Panels (a) and (b) demonstrates the relevance of normal modes, and the underlying mechanical model, to the functioning of adenylate kinase.



(a)



(b)

Figure 6.9: Panel (a): The mean intensity reconstruction of FHV based on 528 cryo-EM experimental images. Panel (b): Mechanical model based on a discretization of the mean intensity shown in (a). Pink spheres represent the point masses and yellow rods represent connecting springs between two neighboring point masses.

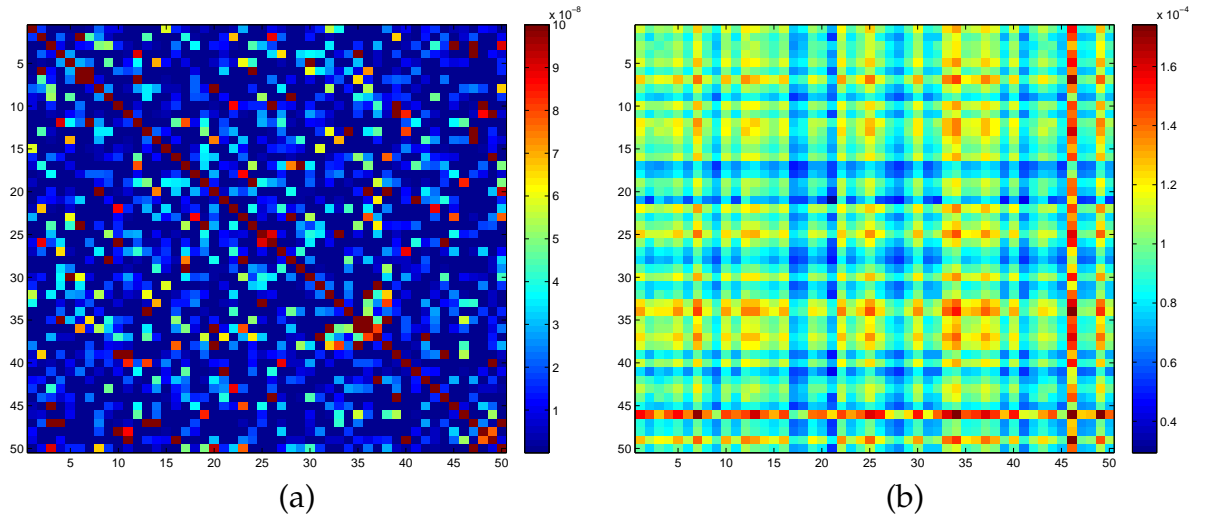


Figure 6.10: Panel (a): Experimental covariance matrix between the 50 point masses from one asymmetric unit of the FHV model. Panel (b): Predicted covariance matrix between the 50 point masses from one asymmetric unit of the FHV model. The spring parameters are the result of the last step in the minimization process, though the predicted mean and covariance never get reasonably close to the experimental mean and covariance.



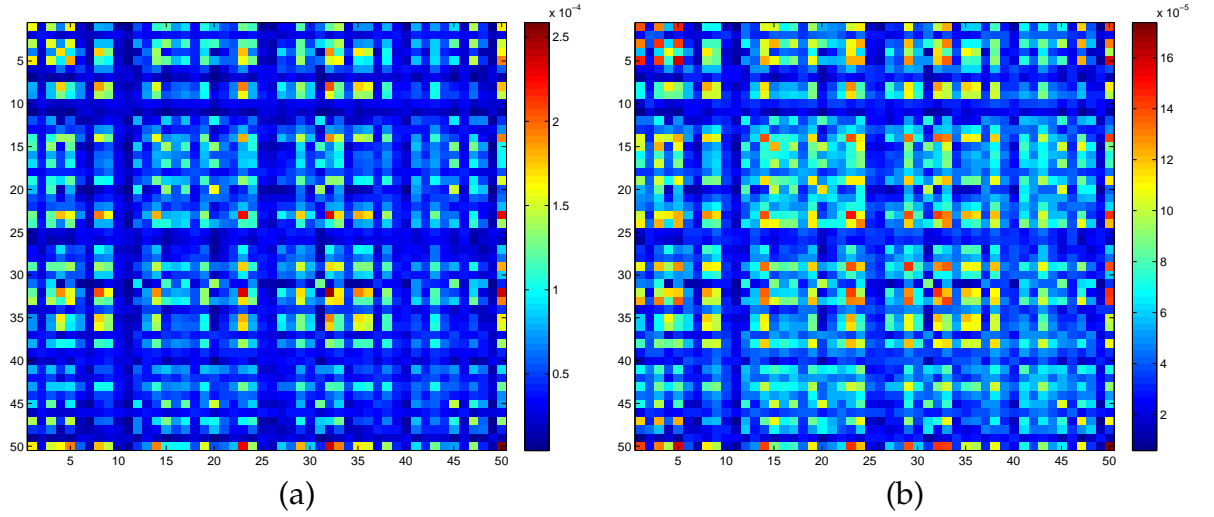


Figure 6.11: A comparison between two Covariance matrix from the FHV model in which the spring constants are randomly chosen between the values of 0 to 100. It can be seen that the overall structure of the covariance matrix is similar even though the actual values might be different. This hints at that the spring connectivity actually places a strong constraint on the Covariance matrix structure. Panel (a): Predicted covariance matrix between the 50 point masses from one asymmetric unit of the FHV model with random spring constant. Panel (b): Predicted covariance matrix between the 50 point masses from one asymmetric unit of the FHV model with another set of random spring constant.

APPENDIX A  
TOMOGRAM PROCESSING ALGORITHMS FOR CHAPTER 3

**A.1 Class Membership, Location, and Orientation Determination of a Particle in a CET Tomogram**

The following tomogram processing routine is based on the maximum likelihood particle structures estimated as in Section 2.6 in order to collect many particles and describe lattices of particles.

1. Using correlation with a spherically-symmetric template and human review of the resulting detections, a set of subcubes is extracted from the tomogram where each subcube contains one particle.
2. Independent of the number of classes of particle that are seen, a one-class reconstruction is performed by the method demonstrated in Section 2.6.
3. By correlating a spherically-averaged version of the one-class reconstruction of Item 2 with the tomogram, the location of each particle in the tomogram is determined.
4. Using the subcubes determined in Item 1, a multiclass reconstruction is performed by the method demonstrated in Section 2.6.
5. For each class in the multiclass reconstruction, a library of 3-D templates is computed where each template is a rotated version of the structure as it would appear in the tomogram for that class. As a part of this calculation, the missing wedge in reciprocal space is removed. The same set of rotations are used as were previously used as the abscissa values in the integral in the expectation step of the expectation maximization algorithm.

6. By correlating the templates of Item 5 with the tomogram at the specific locations determined in Item 3, the class label and the orientation of the particle at each location is determined. The correlation calculation at the  $i$ th location  $\mathbf{x}_i$  is

$$r_{\eta,z}(\mathbf{x}_i) = \frac{\int \rho(\mathbf{x}') t_{\eta,z}(\mathbf{x}' - \mathbf{x}_i) d^3 \mathbf{x}'}{\sqrt{\left[ \int_T \rho^2(\mathbf{x}') d^3 \mathbf{x}' \right] \left[ \int t_{\eta,z}^2(\mathbf{x}') d^3 \mathbf{x}' \right]}} \quad (\text{A.1})$$

where  $T = \text{support } t_{\eta,z}(\mathbf{x}') \subset \mathbf{R}^3$  and  $t_{\eta,z}(\cdot)$  is the template for class  $\eta$  in orientation  $z$  from Item 5. For the  $i$ th particle, the class label  $\eta_i$  and the orientation of the particle  $z_i$  are

$$\eta_i, z_i = \arg \max_{\eta,z} r_{\eta,z}(\mathbf{x}_i). \quad (\text{A.2})$$

The reason for the two-stage approach (Items 3 and 6) is to reduce computation. The result is a list of locations  $\mathbf{x}$ , class  $\eta$ , and orientations  $z$  for all of the particles in the tomogram.

## A.2 Determination of a lattice

1. Determine a cluster of virus particles by manually picking a seed particle and then including in the cluster all particles that are within 60 voxels of the seed particle or of a particle already in the cluster.
2. Manually select a particle in the center of the cluster. Denote its location by  $\mathbf{x}_0$ .
3. Within a sphere of radius  $r_x$ , select the  $n_x$  particles closest to  $\mathbf{x}_0$ . Denote the locations by  $\mathbf{x}_i$  for  $i \in \{1, \dots, n_x\}$ .
4. Compute  $\mathbf{v}_i = \mathbf{x}_i - \mathbf{x}_0$  for  $i \in \{1, \dots, n_x\}$ .

5. Compute the angles between the  $\mathbf{v}_i$  for  $i \in \{1, \dots, n_x\}$  by

$$\theta_{i,j} = \arccos \left( (\mathbf{v}_i / \|\mathbf{v}_i\|)^T (\mathbf{v}_j / \|\mathbf{v}_j\|) \right) \quad (\text{A.3})$$

for  $i, j \in \{1, \dots, n_x\}$

6. In a perfect lattice, particles would occur at locations  $\mathbf{x}_0 + n_a \mathbf{u}_a$  where  $n$  is an integer and similarly for two other directions  $\mathbf{u}_b$  and  $\mathbf{u}_c$  and for mixtures of integer steps in all three directions. The value of  $r_x$  is selected so that only the first particle in any direction, i.e., the  $n_a = 1$  particle in the  $\mathbf{u}_a$  direction, is included among the  $\mathbf{x}_j$ .
7. Among the  $\mathbf{v}_j$  for  $j \in \{1, \dots, n_x\}$ , select  $\mathbf{u}_\iota = \mathbf{v}_{i(\iota)}$  for  $\iota \in \{1, 2, 3\}$  such that  $\theta_{i(\iota), i(\iota')} < \pi/2$  for  $\iota, \iota' \in \{1, 2, 3\}$ . The  $\mathbf{u}_\iota$  are the approximate lattice constants.
8. For each particle (with location  $\mathbf{y}_i$  for  $i \in \{1, \dots, n_y\}$ ), determine its approximate location in terms of the approximate lattice constants by solving the integer optimization problem

$$(h_i, k_i, l_i) = \arg \max_{h,k,l \in \mathcal{Z}} \|\mathbf{y}_i - (\mathbf{x}_0 + h\mathbf{u}_1 + k\mathbf{u}_2 + l\mathbf{u}_3)\|^2 \quad (\text{A.4})$$

where  $\mathcal{Z}$  denotes the integers. The solution of the optimization problem is computed by first computing the solution with  $h_i, k_i, l_i \in \mathbf{R}$ , which is a linear least squares problem, and then rounding the solution to the nearest integers.

9. Using the same integer indices  $(h_i, k_i, l_i)$  for  $i \in \{1, \dots, n_y\}$ , optimize the approximate lattice constants by solving the following optimization problem:

$$(\mathbf{u}_{1,*}, \mathbf{u}_{2,*}, \mathbf{u}_{3,*}) = \arg \min_{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3 \in \mathbf{R}^3} \sum_{i=1}^{n_y} \|\mathbf{y}_i - (\mathbf{x}_0 + h_i \mathbf{u}_1 + k_i \mathbf{u}_2 + l_i \mathbf{u}_3)\|^2 \quad (\text{A.5})$$

where  $\mathbf{R}$  denotes the real numbers.

10. Define  $\delta_i = \mathbf{y}_i - \mathbf{x}_0 \in \mathbb{R}^3$ , and  $\mathbf{n}_i = (h_i, k_i, l_i)^T \in \mathcal{Z}^3$ . Then Eq. A.5 can be written

$$U_* = \arg \max_{U \in \mathbb{R}^{3 \times 3}} \sum_{i=1}^{n_y} \|\delta_i - U \mathbf{n}_i\|^2 \quad (\text{A.6})$$

which is a linear least squares problem for the matrix  $U \in \mathbb{R}^{3 \times 3}$ .

For the calculations on STIV infected *Sulfolobus sulfataricus* cells described in Ref. [35], the parameter values are  $r_x = 61$  voxels and  $n_x = 6$ .

### A.3 Orientational heterogeneity of the particles in a lattice

Let  $\rho(\mathbf{x})$  be the electron scattering intensity function for the particle in the standard orientation. Let  $R_i$  be the rotation matrix that rotates the particle in standard orientation into the orientation taken by the  $i$ th particle in the tomogram, i.e., the  $i$ th particle in the tomogram is a translation of  $\rho_i(\mathbf{x}) = \rho(R_i^{-1} \mathbf{x})$  where  $\rho(\mathbf{x})$  is the particle in standard orientation. Because the particle has icosahedral symmetry, there are 60 equivalent rotation matrices but a unique choice is made in Eq. A.2. Determine the Euler angles  $\alpha_i$ ,  $\beta_i$ , and  $\gamma_i$  that describe the rotation matrix  $R_i$ . Because the coordinate system for the particle in standard orientation has a 5-fold symmetry axis in the  $z$  direction, the pair  $(\alpha_i, \beta_i)$  describe the orientation of a 5-fold symmetry axis. A unit vector along the 5-fold axis is  $\mathbf{u}_i = (\sin \alpha_i \cos \beta_i, \sin \alpha_i \sin \beta_i, \cos \alpha_i)^T$  and the angle between two such unit vectors is  $\theta_{i,i'} = \arccos(\mathbf{u}_i^T \mathbf{u}_{i'})$ . The histogram (not shown) of  $\theta_{i,i'}$  where  $i$  runs over the indices of all particles in the lattice and  $i'(i)$  is the index of the particle closest to  $i$  has multiple peaks indicating that the particle orientations are heterogeneous.

# APPENDIX B

## SPHERICAL FOURIER TRANSFORM AND SO(3) FOURIER TRANSFORM

### B.1 Representing the electron scattering intensity

#### B.1.1 Symbolic

##### Real and reciprocal space in spherical and rectangular coordinates

Let the real-space electron scattering intensity function be denoted by  $\rho(\mathbf{x}) : \mathbb{R}^3 \rightarrow \mathbb{R}$ . Two complete systems of orthonormal basis functions are considered. The first system is spherical harmonics, denoted by  $Y_{l,m}(\theta, \phi) = Y_{l,m}(\mathbf{x}/\|\mathbf{x}\|)$ , which is a system for the surface of the sphere in 3-D, specifically,  $S_2 = \{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\| = 1\}$ . In spherical coordinates, let  $\mathbf{x} = (r, \theta, \phi)$ . Orthonormality is the equation

$$\delta_{l,l'}\delta_{m,m'} = \int_{\Omega} Y_{l,m}(\theta, \phi) Y_{l',m'}^*(\theta, \phi) d\Omega \quad (\text{B.1})$$

$$= \int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} Y_{l,m}(\theta, \phi) Y_{l',m'}^*(\theta, \phi) \sin(\theta) d\theta d\phi. \quad (\text{B.2})$$

Using this system,

$$\rho(\mathbf{x}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} c_{l,m}(r) Y_{l,m}(\theta, \phi) \quad (\text{B.3})$$

$$c_{l,m}(r) = \int_{\Omega} \rho(\mathbf{x}) Y_{l,m}^*(\theta, \phi) d\Omega \quad (\text{B.4})$$

$$= \int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} \rho(\mathbf{x}) Y_{l,m}^*(\theta, \phi) \sin(\theta) d\theta d\phi \quad (\text{B.5})$$

where the weights,  $c_{l,m}(\cdot)$ , depend on the radius  $r$ . This system has two attractive features:

1. When  $\rho(\mathbf{x}) \neq 0$  only in a sphere  $B_3 = \{\mathbf{x} \in \mathbf{R}^3 : \|\mathbf{x}\| \leq 1\}$  it is straightforward to impose the constraint on Eq. B.3 by requiring  $c_{l,m}(r) = 0$  for all  $r > 1$ ,  $l \in \{0, 1, \dots\}$ , and  $m \in \{-l, \dots, +l\}$ . Then,  $c_{l,m}(r)$  can be expanded in a countable collection of basis functions, e.g., certain functions related to spherical Bessel functions [122].
2. Some symmetry constraints, where the symmetries are rotational, can be exactly imposed on the  $c_{l,m}(r)$  coefficients. In particular, the important icosahedral symmetry can be exactly imposed [123, 121].

The second system is complex exponentials, specifically,  $\exp(+i2\pi\mathbf{k}^T\mathbf{x})$ . In rectangular coordinates, let  $\mathbf{x} = (x_1, x_2, x_3)$ . Orthonormality is the equation

$$\delta^{(3)}(\mathbf{k} - \mathbf{k}') = \delta(k_1 - k'_1)\delta(k_2 - k'_2)\delta(k_3 - k'_3) \quad (\text{B.6})$$

$$= \int_{\mathbf{R}^3} \exp(+i2\pi\mathbf{k}^T\mathbf{x}) \left[ \exp(+i2\pi(\mathbf{k}')^T\mathbf{x}) \right]^* d\mathbf{x} \quad (\text{B.7})$$

$$= \int_{\mathbf{R}^3} \exp(+i2\pi(\mathbf{k} - \mathbf{k}')^T\mathbf{x}) d\mathbf{x}. \quad (\text{B.8})$$

Using this system,

$$\rho(\mathbf{x}) = \int_{\mathbf{R}^3} P(\mathbf{k}) \exp(+i2\pi\mathbf{k}^T\mathbf{x}) d\mathbf{k} \quad (\text{B.9})$$

$$P(\mathbf{k}) = \int_{\mathbf{R}^3} \rho(\mathbf{x}) \exp(-i2\pi\mathbf{k}^T\mathbf{x}) d\mathbf{x} \quad (\text{B.10})$$

which is exactly the Fourier Transform pair in  $\mathbf{R}^3$ .

It is desirable to have formulas to interconvert among  $\rho(\mathbf{x})$ ,  $c_{l,m}(r)$ ,  $P(\mathbf{k})$ , and  $C_{l,m}(k)$  (defined in Eq. B.20) as shown in Figure B.1. Equations for the remaining transformations drawn in Figure B.1 are derived in the following paragraphs. In spherical coordinates, in addition to  $\mathbf{x} = (r, \theta, \phi)$ , let  $\mathbf{k} = (k, \theta', \phi')$ . A key tool [53, Eq. 16.127] is

$$\exp(i\mathbf{k}^T\mathbf{x}) = 4\pi \sum_{l=0}^{\infty} i^l j_l(kr) \sum_{m=-l}^{+l} Y_{l,m}^*(\theta, \phi) Y_{l,m}(\theta', \phi') \quad (\text{B.11})$$

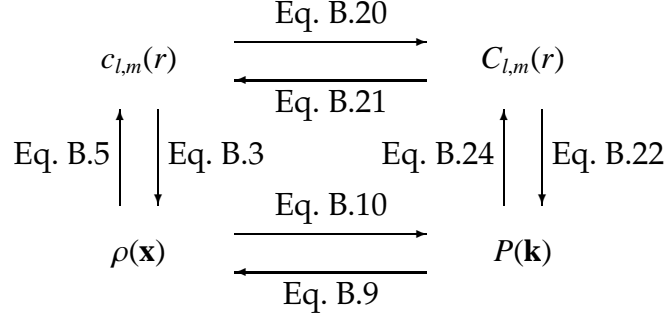


Figure B.1: Conversions among  $\rho(\mathbf{x})$ ,  $c_{l,m}(r)$ ,  $P(\mathbf{k})$ , and  $C_{l,m}(k)$ .

which implies (take complex conjugates and use the fact that  $j_l(\cdot)$  is real)

$$\exp(-i\mathbf{k}^T \mathbf{x}) = 4\pi \sum_{l=0}^{\infty} (-i)^l j_l(kr) \sum_{m=-l}^{+l} Y_{l,m}(\theta, \phi) Y_{l,m}^*(\theta', \phi') \quad (\text{B.12})$$

which implies (interchange  $\mathbf{x}$  and  $\mathbf{k}$  which leaves  $\mathbf{k}^T \mathbf{x}$  unchanged but exchanges  $(\theta, \phi)$  and  $(\theta', \phi')$ )

$$\exp(-i\mathbf{k}^T \mathbf{x}) = 4\pi \sum_{l=0}^{\infty} (-i)^l j_l(kr) \sum_{m=-l}^{+l} Y_{l,m}^*(\theta, \phi) Y_{l,m}(\theta', \phi') \quad (\text{B.13})$$

which implies

$$\exp(-i2\pi\mathbf{k}^T \mathbf{x}) = 4\pi \sum_{l=0}^{\infty} (-i)^l j_l(2\pi kr) \sum_{m=-l}^{+l} Y_{l,m}^*(\theta, \phi) Y_{l,m}(\theta', \phi'). \quad (\text{B.14})$$

Eqs. B.20, B.21, and B.22 are derived in this paragraph. Use Eq. B.3 in Eq. B.10 to get

$$P(\mathbf{k}) = \int_{\mathbf{R}^3} \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} c_{l,m}(r) Y_{l,m}(\theta, \phi) \exp(-i2\pi\mathbf{k}^T \mathbf{x}) d\mathbf{x}. \quad (\text{B.15})$$

Replace  $\exp(-i2\pi\mathbf{k}^T \mathbf{x})$  by Eq. B.14 and do the angular part of  $d\mathbf{x}$  by Eq. B.2 to get

$$P(\mathbf{k}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} (-i)^l \left[ 4\pi \int_{r=0}^{\infty} c_{l,m}(r) j_l(2\pi kr) r^2 dr \right] Y_{l,m}(\theta', \phi'). \quad (\text{B.16})$$

Define the  $l$ th order spherical Hankel transform of a function  $f(r) : \mathbf{R}_+ \rightarrow \mathbf{R}$  by [28, Eq. 7.10.5 (60)][122]

$$F(k) = 4\pi \int_{r=0}^{\infty} f(r) j_l(2\pi kr) r^2 dr \quad (\text{B.17})$$



which has inverse transform

$$f(r) = 4\pi \int_{r=0}^{\infty} F(k) j_l(2\pi kr) k^2 dk. \quad (\text{B.18})$$

The  $l$ th order spherical Hankel transform has a Parseval's theorem, specifically,

$$\int_{r=0}^{\infty} f_1(r) f_2^*(r) r^2 dr = \int_{k=0}^{\infty} F_1(k) F_2^*(k) k^2 dk. \quad (\text{B.19})$$

Applying the  $l$ th order transform to  $c_{l,m}(r)$  gives

$$C_{l,m}(k) = 4\pi \int_{r=0}^{\infty} c_{l,m}(r) j_l(2\pi kr) r^2 dr \quad (\text{B.20})$$

with inverse transform

$$c_{l,m}(r) = 4\pi \int_{k=0}^{\infty} C_{l,m}(k) j_l(2\pi kr) k^2 dk \quad (\text{B.21})$$

and therefore

$$P(\mathbf{k}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} (-i)^l C_{l,m}(k) Y_{l,m}(\theta', \phi'). \quad (\text{B.22})$$

Eq. B.24 is derived in this paragraph. From Eqs. B.2 and B.22 it follows that

$$C_{l,m}(k) = i^l \int_{\Omega'} P(\mathbf{k}) Y_{l,m}^*(\theta', \phi') d\Omega' \quad (\text{B.23})$$

$$= i^l \int_{\theta'=0}^{\pi} \int_{\phi'=0}^{2\pi} P(\mathbf{k}) Y_{l,m}^*(\theta', \phi') \sin(\theta') d\theta' d\phi'. \quad (\text{B.24})$$

Then,  $c_{l,m}(r)$  can be computed from  $C_{l,m}(k)$  by Eq. B.21.

## Rotations

Rotations of  $\mathbf{R}^3$  are described by Euler angles  $(\alpha, \beta, \gamma)$  and by rotation matrices  $R \in \mathbf{R}^{3 \times 3}$  ( $R^{-1} = R^T$ ,  $\det R = +1$ ). Let  $O_R$  be the abstract rotation operator which is defined by

$$\rho'(\mathbf{x}) = O_R\{\rho(\mathbf{x})\} = \rho(R^{-1}\mathbf{x}). \quad (\text{B.25})$$

The Euler angle conventions of Ref. [95] are used and the conversion of Euler angles to rotation matrix is known [95, Eq. 4.43 p. 65]. Let  $R_{\alpha,\beta,\gamma}$  be the rotation matrix corresponding to Euler angles  $(\alpha, \beta, \gamma)$ . Note that  $R_{\alpha,\beta,\gamma}^{-1} = R_{-\gamma, -\beta, -\alpha}$ .

Let  $\rho'(\mathbf{x})$  be defined by Eq. B.25. Then  $P'(\mathbf{k})$  can be evaluated from Eq. B.10 with the result that (change of variables formula and the fact that  $\det R = +1$ )

$$P'(\mathbf{k}) = P(R^{-1}\mathbf{k}) \quad (\text{B.26})$$

where  $P(\mathbf{k})$  corresponds to  $\rho(\mathbf{x})$ .

Spherical harmonics have special properties under rotations. Specifically [95, Eqs. 4.28a and 4.12],

$$O_{R_{\alpha,\beta,\gamma}}\{Y_{l,m}(\theta, \phi)\} = Y_{l,m}(R_{\alpha,\beta,\gamma}^{-1}\mathbf{x}/\|\mathbf{x}\|) \quad (\text{B.27})$$

$$= \sum_{m'=-l}^{+l} D_{m',m}^l(\alpha, \beta, \gamma) Y_{l,m'}(\mathbf{x}/\|\mathbf{x}\|) \quad (\text{B.28})$$

where  $D_{m',m}^l(\alpha, \beta, \gamma)$  are the Wigner  $D$  coefficients [95, Eq. 4.8, p. 52] with values

$$D_{m',m}^l(\alpha, \beta, \gamma) \doteq e^{-im'\alpha} d_{m',m}^l(\beta) e^{-im\gamma} \quad (\text{B.29})$$

and where there are extensive results on the computation of  $d_{m',m}^l(\beta)$  [95, Eq. 4.13], which is real valued.

Suppose that  $\rho'(\mathbf{x})$  is a rotated version of  $\rho(\mathbf{x})$ , specifically,

$$\rho'(\mathbf{x}) = O_{R_{\alpha,\beta,\gamma}}\{\rho(\mathbf{x})\}. \quad (\text{B.30})$$

Then, from Eqs. B.3 and B.28, it follows that

$$\rho'(\mathbf{x}) = O_{R_{\alpha,\beta,\gamma}} \left\{ \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} c_{l,m}(r) Y_{l,m}(\theta, \phi) \right\} \quad (\text{B.31})$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} c_{l,m}(r) O_{R_{\alpha,\beta,\gamma}}\{Y_{l,m}(\theta, \phi)\} \quad (\text{B.32})$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} c_{l,m}(r) \sum_{m'=-l}^{+l} D_{m',m}^l(\alpha, \beta, \gamma) Y_{l,m'}(\mathbf{x}/\|\mathbf{x}\|) \quad (\text{B.33})$$

$$= \sum_{l=0}^{\infty} \sum_{m'=-l}^{+l} \left[ \sum_{m=-l}^{+l} c_{l,m}(r) D_{m',m}^l(\alpha, \beta, \gamma) \right] Y_{l,m'}(\mathbf{x}/\|\mathbf{x}\|) \quad (\text{B.34})$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \left[ \sum_{m'=-l}^{+l} c_{l,m'}(r) D_{m,m'}^l(\alpha, \beta, \gamma) \right] Y_{l,m}(\mathbf{x}/\|\mathbf{x}\|) \quad (\text{B.35})$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} c'_{l,m}(r) Y_{l,m}(\mathbf{x}/\|\mathbf{x}\|) \quad (\text{B.36})$$

where

$$c'_{l,m}(r) = \sum_{m'=-l}^{+l} c_{l,m'}(r) D_{m,m'}^l(\alpha, \beta, \gamma). \quad (\text{B.37})$$

Eq. B.37 implies

$$C'_{l,m}(k) = \sum_{m'=-l}^{+l} C_{l,m'}(k) D_{m,m'}^l(\alpha, \beta, \gamma). \quad (\text{B.38})$$

Notice that the  $m$  and  $m'$  indices of the Wigner  $D$  coefficient in Eqs. B.37 and B.38 are reversed relative to the indices in Eq. B.28.

In summary, Eqs. B.25, B.26, B.37, and B.38 describe how the four key quantities,  $\rho(\mathbf{x})$ ,  $P(\mathbf{k})$ ,  $c_{l,m}(r)$ , and  $C_{l,m}(k)$ , respectively, transform under a rotation.

### B.1.2 Computation

A key idea is to compute collectively for all possible values of the independent variable. This is not a novel idea. For instance, suppose there are two real-valued discrete-time signals, denoted by  $x[n]$  and  $e[n]$ , which are defined only for  $n \in \{0, 1, \dots, N-1\}$ . Define

$$x^\infty[n] = \begin{cases} x[n], & n \in \{0, 1, \dots, N-1\} \\ 0, & \text{otherwise} \end{cases} \quad (\text{B.39})$$

$$e^\infty[n] = \begin{cases} e[n], & n \in \{0, 1, \dots, N-1\} \\ 0, & \text{otherwise} \end{cases}. \quad (\text{B.40})$$

The goal is to compute

$$r^\infty[n] = \sum_{m=-\infty}^{+\infty} x^\infty[m]e^\infty[m-n]. \quad (\text{B.41})$$

An efficient approach, assuming that you desire all nonzero values of the sequence  $r^\infty[\cdot]$ , is to define  $L = 2N + 1$ , to define

$$x_L[n] = \begin{cases} x[n], & n \in \{0, 1, \dots, N-1\} \\ 0, & n \in \{N, \dots, L-1\} \end{cases} \quad (\text{B.42})$$

$$e_L[n] = \begin{cases} e[n], & n \in \{0, 1, \dots, N-1\} \\ 0, & n \in \{N, \dots, L-1\} \end{cases}, \quad (\text{B.43})$$

to take the  $L$ -point FFT of both  $x_L[\cdot]$  and of  $e_L[\cdot]$  (denoted by  $X_L[\cdot]$  and  $E_L[\cdot]$ , respectively), to compute  $R_L[k] = X_L[k]E_L^*[k]$ , and finally to compute the inverse FFT of  $R_L[\cdot]$  to compute all of the nonzero values of  $r^\infty[\cdot]$ .

## Real and reciprocal space in spherical and rectangular coordinates

Figure B.1 shows four quantities and eight conversions. But there are really more quantities since  $\rho(\mathbf{x})$  and  $P(\mathbf{k})$  must be sampled and different sampling schemes are natural for different computations.

We consider two sampling schemes:

### 1. Rectangular sampling:

$$\mathbf{x}_{\mathbf{n}} = \begin{bmatrix} \delta_{x_1} n_1 \\ \delta_{x_2} n_2 \\ \delta_{x_3} n_3 \end{bmatrix}, \quad (\mathbf{n} \in \mathcal{Z}^3) \quad (\text{B.44})$$

and similarly for  $\mathbf{k}_n$ .

## 2. Spherical sampling:

$$\mathbf{x}_{n_r, n_\theta, n_\phi} = (\delta_r n_r, \text{something for } \theta, \text{something for } \phi), \quad (n_r \in \mathcal{Z}_+ \cup \{0\}, \text{more for } \theta \text{ and } \phi) \quad (\text{B.45})$$

and similarly for  $\mathbf{k}_{n_k, n_{\theta'}, n_{\phi'}}$ .

We assume  $\rho(\mathbf{x})$  is smooth so conversion between rectangular and spherical sampling schemes can be done by simple interpolators, and we have used trilinear interpolation. However,  $P(\mathbf{k})$  is not smooth and so simple interpolators are problematic. Therefore, we avoid conversion between rectangular and spherical sampling schemes for  $P(\mathbf{k})$ .

We have used the following algorithms, which are not a complete set:

1.  $\rho(\mathbf{x})$  (rectangular sampling) to  $P(\mathbf{k})$  (rectangular sampling): 3-D FFT.
2.  $P(\mathbf{k})$  (rectangular sampling) to  $\rho(\mathbf{x})$  (rectangular sampling): 3-D inverse FFT.
3.  $\rho(\mathbf{x})$  (rectangular or spherical sampling) to  $c_{l,m}(r)$  (orthonormal expansion): numerical quadrature approximation of Eq. B.5 or least squares.
4.  $c_{l,m}(r)$  (sampled, finite support) to  $C_{l,m}(k)$  (sampled): numerical quadrature of the  $l$ th order spherical Hankel transform integral with a truncated range for  $k$ .
5.  $c_{l,m}(r)$  (orthonormal expansion) to  $C_{l,m}(k)$  (orthonormal expansion) and *visa versa*: Direct evaluation of Eqs. B.77 and B.78.

## Rotations

We have used the following algorithms, which are not a complete set:

1. One particular rotation of  $c_{l,m}(r)$  or  $C_{l,m}(k)$ : Direct evaluation of Eq. B.37 or B.38 using Algorithm 1.

---

Algorithm 1: One specific rotation of  $c_{l,m}(r)$  or  $C_{l,m}(k)$  described by  $(\alpha, \beta, \gamma)$ . The case of  $c_{l,m}(r)$  and  $c'_{l,m}(r)$  is described. The case of  $C_{l,m}(k)$  and  $C'_{l,m}(k)$  is the same.

```
for  $l = 1 \rightarrow \infty$  do  
  compute  $D_{\cdot,\cdot}^l(\alpha, \beta, \gamma)$  as a  $2l+1 \times 2l+1$  matrix using WHAT PART OF DART-  
  MOUTH?  
  for  $r = R_1 \rightarrow R_2$  do  
     $c'_{l,\cdot}(r) = D_{\cdot,\cdot}^l(\alpha, \beta, \gamma)c_{l,\cdot}(r)$   
  end for  
end for
```

---

2. One particular rotation of  $\rho(\mathbf{x})$ : Approximate evaluation of Eq. B.25 using trilinear interpolation.

## B.2 Correlation

### B.2.1 Symbolic

The correlation of two signals is defined by

$$R_{\rho^{(1)}, \rho^{(2)}} = \int_{\mathbf{R}^3} \rho^{(1)}(\mathbf{x}) [\rho^{(2)}(\mathbf{x})]^* d\mathbf{x}. \quad (\text{B.46})$$

By Parseval's theorem,

$$R_{\rho^{(1)}, \rho^{(2)}} = \int_{\mathbf{R}^3} \rho^{(1)}(\mathbf{x}) [\rho^{(2)}(\mathbf{x})]^* d\mathbf{x} = \int_{\mathbf{R}^3} P^{(1)}(\mathbf{k}) [P^{(2)}(\mathbf{k})]^* d\mathbf{k}. \quad (\text{B.47})$$

By Eq. B.3 and then Eq. B.2 it follows that

$$R_{\rho^{(1)}, \rho^{(2)}} = \int_{\mathbf{R}^3} \left[ \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} c_{l,m}^{(1)}(r) Y_{l,m}(\theta, \phi) \right] \left[ \sum_{l'=0}^{\infty} \sum_{m'=-l'}^{+l'} c_{l',m'}^{(2)}(r) Y_{l',m'}(\theta, \phi) \right]^* d\mathbf{x} \quad (\text{B.48})$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \int_{r=0}^{\infty} c_{l,m}^{(1)}(r) [c_{l,m}^{(2)}(r)]^* r^2 dr. \quad (\text{B.49})$$

By using Eqs. B.47, B.22, and B.2, it follows that

$$R_{\rho^{(1)}, \rho^{(2)}} = \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \int_{k=0}^{\infty} C_{l,m}^{(1)}(k) [C_{l,m}^{(2)}(k)]^* k^2 dk. \quad (\text{B.50})$$

In summary, correlation can be computed from  $\rho(\mathbf{x})$ ,  $P(\mathbf{k})$ ,  $c_{l,m}(r)$ , or  $C_{l,m}(k)$  by using Eqs. B.46, B.47, B.49, or B.50, respectively:

$$R_{\rho^{(1)}, \rho^{(2)}} = \int_{\mathbf{R}^3} \rho^{(1)}(\mathbf{x}) [\rho^{(2)}(\mathbf{x})]^* d\mathbf{x} \quad (\text{B.51})$$

$$= \int_{\mathbf{R}^3} P^{(1)}(\mathbf{k}) [P^{(2)}(\mathbf{k})]^* d\mathbf{k} \quad (\text{B.52})$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \int_{r=0}^{\infty} c_{l,m}^{(1)}(r) [c_{l,m}^{(2)}(r)]^* r^2 dr \quad (\text{B.53})$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \int_{k=0}^{\infty} C_{l,m}^{(1)}(k) [C_{l,m}^{(2)}(k)]^* k^2 dk. \quad (\text{B.54})$$

Often it is desired to compute correlation for all rotations of  $\rho^{(2)}(\mathbf{x})$ . By using Eqs. B.25, B.26, B.37, and B.38 in Eqs. B.46, B.47, B.49, or B.50, respectively, it follows that

$$\begin{aligned} R_{\rho^{(1)}, \rho^{(2)}}(\alpha, \beta, \gamma) \\ = \int_{\mathbf{R}^3} \rho^{(1)}(\mathbf{x}) [\rho^{(2)}(R_{\alpha, \beta, \gamma}^{-1} \mathbf{x})]^* d\mathbf{x} \end{aligned} \quad (\text{B.55})$$

$$= \int_{\mathbf{R}^3} P^{(1)}(\mathbf{k}) [P^{(2)}(R_{\alpha,\beta,\gamma}^{-1} \mathbf{k})]^* d\mathbf{k} \quad (\text{B.56})$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \int_{r=0}^{\infty} c_{l,m}^{(1)}(r) \left[ \sum_{m'=-l}^{+l} c_{l,m'}^{(2)}(r) D_{m,m'}^l(\alpha, \beta, \gamma) \right]^* r^2 dr \quad (\text{B.57})$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \sum_{m'=-l}^{+l} \left[ \int_{r=0}^{\infty} c_{l,m}^{(1)}(r) [c_{l,m'}^{(2)}(r)]^* r^2 dr \right] [D_{m,m'}^l(\alpha, \beta, \gamma)]^* \quad (\text{B.58})$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \sum_{m'=-l}^{+l} \omega_{m,m'}^l [D_{m,m'}^l(\alpha, \beta, \gamma)]^* \quad (\text{B.59})$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \int_{k=0}^{\infty} C_{l,m}^{(1)}(k) \left[ \sum_{m'=-l}^{+l} C_{l,m'}^{(2)}(k) D_{m,m'}^l(\alpha, \beta, \gamma) \right]^* k^2 dk \quad (\text{B.60})$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \sum_{m'=-l}^{+l} \left[ \int_{k=0}^{\infty} C_{l,m}^{(1)}(k) [C_{l,m'}^{(2)}(k)]^* k^2 dk \right] [D_{m,m'}^l(\alpha, \beta, \gamma)]^* \quad (\text{B.61})$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \sum_{m'=-l}^{+l} \Omega_{m,m'}^l [D_{m,m'}^l(\alpha, \beta, \gamma)]^* \quad (\text{B.62})$$

where

$$\omega_{m,m'}^l = \int_{r=0}^{\infty} c_{l,m}^{(1)}(r) [c_{l,m'}^{(2)}(r)]^* r^2 dr \quad (\text{B.63})$$

$$\Omega_{m,m'}^l = \int_{k=0}^{\infty} C_{l,m}^{(1)}(k) [C_{l,m'}^{(2)}(k)]^* k^2 dk. \quad (\text{B.64})$$

From Parseval's theorem for  $l$ th-order spherical Hankel transforms (Eq. B.19) and the fact that both factors in the integrands of Eqs. B.63 and B.64 share the same value of  $l$  it follows that

$$\omega_{m,m'}^l = \Omega_{m,m'}^l. \quad (\text{B.65})$$

From Eq. B.29 and the fact that  $d_{m',m}^l(\beta) \in \mathbf{R}$  [95, Eq. 4.13], it follows that

$$[D_{m',m}^l(\alpha, \beta, \gamma)]^* = D_{-m',-m}^l(\alpha, \beta, \gamma). \quad (\text{B.66})$$

Therefore,

$$\begin{aligned} & R_{\rho^{(1)}, \rho^{(2)}}(\alpha, \beta, \gamma) \\ &= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \sum_{m'=-l}^{+l} \omega_{-m,-m'}^l D_{m,m'}^l(\alpha, \beta, \gamma) \end{aligned} \quad (\text{B.67})$$



$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \sum_{m'=-l}^{+l} \Omega_{-m,-m'}^l D_{m,m'}^l(\alpha, \beta, \gamma). \quad (\text{B.68})$$

Computing  $\omega_{m,m'}^l$  or  $\Omega_{-m,-m'}^l$  first and then computing the triple sum over  $(l, m, m')$  has lower computational cost than the reverse ordering of the operations.

Eqs. B.67 and B.68 are special cases of the fact that the Wigner  $D$  coefficients, considered as a set of triply-indexed functions of the Euler angles, are a complete system of orthogonal (not orthonormal) basis functions for smooth functions of rotations (here described by Euler angles). The orthogonality (not orthonormality) relationship is the equation [63]

$$\frac{8\pi^2}{2l+1} \delta_{m,\mu} \delta_{m',\mu'} \delta_{l,\lambda} = \int_{\alpha=0}^{2\pi} d\alpha \int_{\beta=0}^{\pi} \sin(\beta) d\beta \int_{\gamma=0}^{2\pi} d\gamma D_{m',m}^l(\alpha, \beta, \gamma) [D_{\mu',\mu}^l(\alpha, \beta, \gamma)]^*. \quad (\text{B.69})$$

Therefore, any smooth function  $f(\alpha, \beta, \gamma)$  can be expressed in the form

$$f(\alpha, \beta, \gamma) = \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \sum_{m'=-l}^{+l} w_{m,m'}^l D_{m,m'}^l(\alpha, \beta, \gamma) \quad (\text{B.70})$$

where

$$w_{m,m'}^l = \frac{2l+1}{8\pi^2} \int_{\alpha=0}^{2\pi} d\alpha \int_{\beta=0}^{\pi} \sin(\beta) d\beta \int_{\gamma=0}^{2\pi} d\gamma f(\alpha, \beta, \gamma) D_{m,m'}^l(\alpha, \beta, \gamma). \quad (\text{B.71})$$

From this point of view, the key contribution of Eqs. B.67 and B.68 is to specify the weights  $w_{m,m'}^l$ .

## B.2.2 Computation

There does not appear to be an efficient approach to computing correlation for all rotations of  $\rho^{(2)}(\mathbf{x})$  based on  $\rho(\mathbf{x})$  or  $P(\mathbf{k})$  (i.e., on Eqs. B.55 or B.56).

Efficient methods to compute correlation for all rotations of  $\rho^{(2)}(\mathbf{x})$  based on  $c_{l,m}(r)$  or  $C_{l,m}(k)$  (i.e., on Eqs. B.67 and B.68) can be based on fast algorithms for

the  $SO(3)$  transform. Because  $\omega_{m,m'}^l = \Omega_{m,m'}^l$  (Eq. B.65), there is little difference between basing the calculation on  $c_{l,m}(r)$  or  $C_{l,m}(k)$  (Eqs. B.67 and B.68). If  $\rho(\mathbf{x})$  has finite support then use  $c_{l,m}(r)$  while if  $\rho(\mathbf{x})$  has finite bandwidth (i.e.,  $P(\mathbf{k})$  has finite support) then use  $C_{l,m}(k)$ .

### B.3 Representing $c_{l,m}(r)$ and $C_{l,m}(k)$ for computation

Two approaches for representing  $c_{l,m}(r)$  and  $C_{l,m}(k)$  for computation are considered. The first approach is periodic sampling. Specifically,

$$c_{l,m}[n] = c_{l,m}(\delta_r n), \quad (n \in \mathcal{Z}) \quad (\text{B.72})$$

$$C_{l,m}[n] = C_{l,m}(\delta_k n), \quad (n \in \mathcal{Z}) \quad (\text{B.73})$$

$$\delta_k = \text{what is the relationship to } \delta_r?. \quad (\text{B.74})$$

If the support of  $\rho(\mathbf{x})$  is finite, which is the usual case, then the support of  $c_{l,m}(r)$  is correspondingly finite and the computation of  $C_{l,m}(k)$  from  $c_{l,m}(r)$  (Hankel transform, Eq. B.20) and  $\omega_{m,m'}^l$  from  $c_{l,m}(r)$  (pointwise multiplication, Eq. B.63) by numerical quadrature (we have used a trapezoidal rule) is straightforward. However, the computation of  $c_{l,m}(r)$  from  $C_{l,m}(k)$  by numerical quadrature (Hankel transform, Eq. B.21) is difficult because the support of  $C_{l,m}(k)$  will typically be all of  $\mathbb{R}_+$ . If the support of  $P(\mathbf{k})$  is finite (i.e.,  $\rho(\mathbf{x})$  is bandlimited), then the support of  $C_{l,m}(k)$  is correspondingly finite and the computation of  $c_{l,m}(r)$  from  $C_{l,m}(k)$  (Hankel transform, Eq. B.21) and  $\Omega_{m,m'}^l$  from  $C_{l,m}(k)$  (pointwise multiplication, Eq. B.64) by numerical quadrature (we have used a trapezoidal rule) is straightforward. However, the computation of  $C_{l,m}(k)$  from  $c_{l,m}(r)$  by numerical quadrature (Hankel transform, Eq. B.20) is difficult because the support of  $c_{l,m}(r)$  will typically be all of  $\mathbb{R}_+$ .

The second approach is to represent  $c_{l,m}(r)$  and  $C_{l,m}(k)$  by orthonormal expansions, i.e.,

$$c_{l,m}(r) = \sum_{p=1}^{\infty} c_{l,m,p} \phi_p(r) \quad (\text{B.75})$$

$$C_{l,m}(k) = \sum_{p=1}^{\infty} C_{l,m,p} \Phi_p(k) \quad (\text{B.76})$$

where  $\phi_p(r)$  and  $\Phi_p(k)$  are complete systems of orthonormal basis functions on  $R_+$  or an appropriate subset of  $R_+$ . Based on Eqs. B.20 and B.21, if  $\phi_p(r)$  and  $\Phi_p(k)$  are spherical Hankel transforms of each other then  $c_{l,m,p} = C_{l,m,p}$ . Since it is necessary to consider spherical Hankel transforms of all orders  $l$ , it is natural to let the basis functions also depend on  $l$ . Sturm-Liouville theory for  $l$ th-order spherical Bessel functions is a natural method to generate complete systems of orthonormal basis functions in real space for the regions  $0 \leq R$  and  $R_1 \leq r \leq R_2$  where both the basis functions (denoted by  $h_{l,p}(r)$ ) and their  $l$ th-order spherical Hankel transforms (denoted by  $H_{l,p}(k)$ ) have explicit symbolic formulas [122]. The result are the representations

$$c_{l,m}(r) = \sum_{p=1}^{\infty} c_{l,m,p} h_{l,p}(r) \quad (\text{B.77})$$

$$C_{l,m}(k) = \sum_{p=1}^{\infty} c_{l,m,p} H_{l,p}(k). \quad (\text{B.78})$$

Parseval's theorem for  $l$ th-order spherical Hankel transforms implies that if  $h_{l,m}(r)$  (for  $m$  varying) is an orthonormal system then  $H_{l,m}(r)$  (for  $m$  varying) is also an orthonormal system.

For the second approach, there are simple formulas for  $\omega_{m,m'}^l$  (Eq. B.63) and  $\Omega_{m,m'}^l$  (Eq. B.64). In particular,

$$\omega_{m,m'}^l = \int_{r=0}^{\infty} c_{l,m}^{(1)}(r) [c_{l,m'}^{(2)}(r)]^* r^2 dr \quad (\text{B.79})$$

$$= \int_{r=0}^{\infty} \left[ \sum_{p=1}^{\infty} c_{l,m,p}^{(1)} h_{l,p}(r) \right] \left[ \sum_{p'=1}^{\infty} c_{l,m',p'}^{(2)} h_{l,p'}(r) \right]^* r^2 dr \quad (\text{B.80})$$

$$= \sum_{p=1}^{\infty} \sum_{p'=1}^{\infty} \left[ \int_{r=0}^{\infty} h_{l,p}(r) [h_{l,p'}(r)]^* r^2 dr \right] c_{l,m,p}^{(1)} [c_{l,m',p'}^{(2)}]^* \quad (\text{B.81})$$

$$= \sum_{p=1}^{\infty} \sum_{p'=1}^{\infty} \delta_{p,p'} c_{l,m,p}^{(1)} [c_{l,m',p'}^{(2)}]^* \quad (\text{B.82})$$

$$= \sum_{p=1}^{\infty} c_{l,m,p}^{(1)} [c_{l,m',p}^{(2)}]^* \quad (\text{B.83})$$

and likewise

$$\Omega_{m,m'}^l = \int_{k=0}^{\infty} C_{l,m}^{(1)}(k) [C_{l,m'}^{(2)}(k)]^* k^2 dk \quad (\text{B.84})$$

$$= \sum_{p=1}^{\infty} c_{l,m,p}^{(1)} [c_{l,m',p}^{(2)}]^* \quad (\text{B.85})$$

which demonstrates explicitly that  $\omega_{m,m'}^l = \Omega_{m,m'}^l$  which was previously demonstrated (Eq. B.65) by Parseval's theorem for  $l$ th-order spherical Hankel transforms (Eq. B.19).

## B.4 Quantities related to correlation

Computing correlation (Eq. B.46) requires substantial computation. Also, if the SNR is low, some authors have believed that additional averaging before computing correlation is helpful [8].

Applying the Cauchy-Schwartz inequality to the definition of correlation (Eq. B.46) gives

$$|R_{\rho^{(1)}, \rho^{(2)}}| = \left| \int_{\mathbf{R}^3} \rho^{(1)}(\mathbf{x}) [\rho^{(2)}(\mathbf{x})]^* d\mathbf{x} \right| \quad (\text{B.86})$$

$$= \left| \int_{\Omega} \int_{r=0}^{\infty} \rho^{(1)}(\mathbf{x}) [\rho^{(2)}(\mathbf{x})]^* r^2 dr d\Omega \right| \quad (\text{B.87})$$

$$\leq \int_{\Omega} \left| \int_{r=0}^{\infty} \rho^{(1)}(\mathbf{x}) [\rho^{(2)}(\mathbf{x})]^* r^2 dr \right| d\Omega \quad (\text{B.88})$$

$$\leq \int_{\Omega} \sqrt{\left[ \int_{r=0}^{\infty} |\rho^{(1)}(\mathbf{x})|^2 r^2 dr \right] \left[ \int_{r=0}^{\infty} |\rho^{(2)}(\mathbf{x})|^2 r^2 dr \right]} d\Omega \quad (\text{B.89})$$

where the function space is the  $r^2$  weighted  $L_p$  space. [12] Similar calculations can be done for Eqs. B.47, B.49, or B.50. In summary of all of these calculations,

$$|R_{\rho^{(1)}, \rho^{(2)}}| \leq \int_{\Omega} \sqrt{\left[ \int_{r=0}^{\infty} |\rho^{(1)}(\mathbf{x})|^2 r^2 dr \right] \left[ \int_{r=0}^{\infty} |\rho^{(2)}(\mathbf{x})|^2 r^2 dr \right]} d\Omega \quad (\text{B.90})$$

$$|R_{\rho^{(1)}, \rho^{(2)}}| \leq \int_{\Omega'} \sqrt{\left[ \int_{k=0}^{\infty} |P^{(1)}(\mathbf{x})|^2 k^2 dk \right] \left[ \int_{k=0}^{\infty} |P^{(2)}(\mathbf{x})|^2 k^2 dk \right]} d\Omega' \quad (\text{B.91})$$

$$|R_{\rho^{(1)}, \rho^{(2)}}| \leq \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \sqrt{\left[ \int_{r=0}^{\infty} |c_{l,m}^{(1)}(r)|^2 r^2 dr \right] \left[ \int_{r=0}^{\infty} |c_{l,m}^{(2)}(r)|^2 r^2 dr \right]} \quad (\text{B.92})$$

$$|R_{\rho^{(1)}, \rho^{(2)}}| \leq \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \sqrt{\left[ \int_{k=0}^{\infty} |C_{l,m}^{(1)}(k)|^2 k^2 dk \right] \left[ \int_{k=0}^{\infty} |C_{l,m}^{(2)}(k)|^2 k^2 dk \right]}. \quad (\text{B.93})$$

If an orthonormal expansion is used to represent  $c_{l,m}^{(j)}(r)$  and  $C_{l,m}^{(1)}(k)$  functions (Eqs. B.77 and B.78) then

$$\begin{aligned} & \int_{r=0}^{\infty} |c_{l,m}(r)|^2 r^2 dr \\ &= \int_{r=0}^{\infty} \left[ \sum_{p=1}^{\infty} c_{l,m,p} h_{l,p}(r) \right] \left[ \sum_{p'=1}^{\infty} c_{l,m,p'} h_{l,p'}(r) \right]^* r^2 dr \end{aligned} \quad (\text{B.94})$$

$$= \sum_{p=1}^{\infty} \sum_{p'=1}^{\infty} c_{l,m,p} c_{l,m,p'}^* \int_{r=0}^{\infty} h_{l,p}(r) h_{l,p'}^*(r) r^2 dr \quad (\text{B.95})$$

$$= \sum_{p=1}^{\infty} \sum_{p'=1}^{\infty} c_{l,m,p} c_{l,m,p'}^* \delta_{p,p'} \quad (\text{B.96})$$

$$= \sum_{p=1}^{\infty} |c_{l,m,p}|^2 \quad (\text{B.97})$$

for either the first or second function and similarly for either the first or the second  $C_{l,m}(k)$  function. Therefore, Eqs. B.92 and B.93 are equivalent to

$$|R_{\rho^{(1)}, \rho^{(2)}}| \leq \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \sqrt{\left[ \sum_{p=1}^{\infty} |c_{l,m,p}|^2 \right] \left[ \sum_{p=1}^{\infty} |c_{l,m,p}|^2 \right]}. \quad (\text{B.98})$$

When it is desired to compute correlation for all rotations of  $\rho^{(2)}(\mathbf{x})$ , then Eqs. B.55, B.56, B.59, and B.62 can be bounded using the same method to find that

$$|R_{\rho^{(1)}, \rho^{(2)}}(\alpha, \beta, \gamma)| \leq \int_{\Omega} \sqrt{s^{(1)}(\mathbf{x}/\|\mathbf{x}\|) s^{(1)}(R_{\alpha, \beta, \gamma}^{-1} \mathbf{x}/\|\mathbf{x}\|)} d\Omega \quad (\text{B.99})$$

$$s^{(1)}(\theta, \phi) \doteq \int_{r=0}^{\infty} |\rho^{(1)}(\mathbf{x})|^2 r^2 dr \quad (\text{B.100})$$

$$s^{(2)}(\theta, \phi) \doteq \int_{r=0}^{\infty} |\rho^{(2)}(\mathbf{x})|^2 r^2 dr \quad (\text{B.101})$$

$$|R_{\rho^{(1)}, \rho^{(2)}}(\alpha, \beta, \gamma)| \leq \int_{\Omega} \sqrt{S^{(1)}(\mathbf{x}/\|\mathbf{x}\|) S^{(1)}(R_{\alpha, \beta, \gamma}^{-1} \mathbf{x}/\|\mathbf{x}\|)} d\Omega \quad (\text{B.102})$$

$$S^{(1)}(\theta', \phi') \doteq \int_{k=0}^{\infty} |P^{(1)}(\mathbf{x})|^2 k^2 dk \quad (\text{B.103})$$

$$S^{(2)}(\theta', \phi') \doteq \int_{k=0}^{\infty} |P^{(2)}(\mathbf{x})|^2 k^2 dk \quad (\text{B.104})$$

$$|R_{\rho^{(1)}, \rho^{(2)}}(\alpha, \beta, \gamma)| \leq \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \sum_{m'=-l}^{+l} |\omega_{m, m'}^l| |D_{m, m'}^l(\alpha, \beta, \gamma)| \quad (\text{B.105})$$

$$|R_{\rho^{(1)}, \rho^{(2)}}(\alpha, \beta, \gamma)| \leq \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \sum_{m'=-l}^{+l} |\Omega_{m, m'}^l| |D_{m, m'}^l(\alpha, \beta, \gamma)| \quad (\text{B.106})$$

$$|\omega_{m, m'}^l| \leq \sqrt{\left[ \int_{r=0}^{\infty} |c_{l, m}^{(1)}(r)|^2 r^2 dr \right] \left[ \int_{r=0}^{\infty} |c_{l, m'}^{(2)}(r)|^2 r^2 dr \right]} \quad (\text{B.107})$$

$$|\Omega_{m, m'}^l| \leq \sqrt{\left[ \int_{k=0}^{\infty} |C_{l, m}^{(1)}(k)|^2 k^2 dk \right] \left[ \int_{k=0}^{\infty} |C_{l, m'}^{(2)}(k)|^2 k^2 dk \right]} \quad (\text{B.108})$$

where  $s^{(1)}(\theta, \phi)$ ,  $s^{(2)}(\theta, \phi)$ ,  $S^{(1)}(\theta', \phi')$ , and  $S^{(2)}(\theta', \phi')$  are used to emphasize that the radial integrals (in  $r$  or in  $k$ ) can be done before any rotations.

If an orthonormal expansion is used to represent  $c_{l, m}^{(j)}(r)$  and  $C_{l, m}^{(1)}(k)$  functions (Eqs. B.77 and B.78) then  $\omega_{m, m'}^l$  and  $\Omega_{m, m'}^l$  are given by Eqs. B.83 and B.85, respectively. It is natural use the same Cauchy-Schwartz bound on  $\omega_{m, m'}^l$  and  $\Omega_{m, m'}^l$  which leads to the result that

$$|\omega_{m, m'}^l| = |\Omega_{m, m'}^l| \quad (\text{B.109})$$

$$= \left| \sum_{p=1}^{\infty} c_{l, m, p}^{(1)} [c_{l, m', p}^{(2)}]^* \right| \quad (\text{B.110})$$

$$\leq \sqrt{\left[\sum_{p=1}^{\infty} |c_{l,m,p}^{(1)}|^2\right] \left[\sum_{p=1}^{\infty} |c_{l,m',p}^{(2)}|^2\right]} \quad (\text{B.111})$$

which can be combined with Eqs. B.105 or B.106 to provide a bound on  $|R_{\rho^{(1)},\rho^{(2)}}(\alpha,\beta,\gamma)|$ .

## BIBLIOGRAPHY

- [1] N.G. Abrescia, J.J. Cockburn, J.M. Grimes, G.C. Sutton, J.M. Diprose, S.J. Butcher, S.D. Fuller, San Martin, R.M. C., Burnett, and et al. Stuart, D.I. Insights into assembly from structural analysis of bacteriophage PRD1. *Science*, 432:68–74, 2004.
- [2] N.G. Abrescia, J.M. Grimes, H.M. Kivela, R. Assenberg, G.C. Sutton, S.J. Butcher, J.K. Bamford, D.H. Bamford, and D.I. Stuart. Insights into virus evolution and membrane biogenesis from the structure of the marine lipid-containing bacteriophage PM2. *Mol. Cell*, 31:749–761, 2008.
- [3] A. Al-Amoudi, L.P. Norlen, and J. Dubochet. Cryo-electron microscopy of vitreous sections of native biological cells and tissues. *J. Struct. Biol.*, 148:131–135, 2004.
- [4] Frank Alber, Friedrich Forster, Dmitry Korkin, Maya Topf, and Andrej Sali. Integrating diverse data for structure determination of macromolecular assemblies. *Annual Review of Biochemistry*, 77:443–477, 2008.
- [5] Gonzalo Alvarez, Phani K. V. V. Nukala, and Eduardo D’Azevedo. Fast diagonalization of evolving matrices: application to spin-fermion models. *Journal of Statistical Mechanics: Theory and Experiment*, 2007:P08007, 2007.
- [6] Franz Aurenhammer. Voronoi diagrams—A survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3):345–405, 1991.
- [7] D.H. Bamford, R.M. Burnett, and D.I. Stuart. Evolution of viral structure. *Theor. Popul. Biol.*, 61:461–470, 2002.
- [8] A. Bartesaghi, P. Sprechmann, J. Liu, G. Randall, G. Sapiro, and S. Subramaniam. Classification and 3D averaging with missing wedge correction in biological electron tomography. *Journal of Structural Biology*, 162:436–450, June 2008.
- [9] ke Björck. *Numerical Methods for Least Squares Problems*. SIAM, 1996.
- [10] J. Bohm, A.S. Frangakis, R. Hegerl, S. Nickell, D. Typke, and W. Baumeister. Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proc. Nat. Acad. Sci. U.S.A.*, 97:14245–14250, 2000.



- [11] L. Bongini, D. Fanelli, P. De Los Rios, and U. Skoglund. Freezing immunoglobulins to see them move. *Proc. Nat. Acad. Sci. U.S.A.*, 101:6466–6471, 2004.
- [12] Nicolas Bourbaki. *Topological vector spaces, Elements of mathematics*. Springer-Verlag, 1987.
- [13] A. Briegel, D. P. Dias, Z. Li, R. B. Jensen, A. S. Frangakis, and G. J. Jensen. Multiple large filament bundles observed in *caulobacter crescentus* by electron cryotomography. *Mol. Microbiol.*, 62:5–14, 2006.
- [14] Jacob Brink, Steven J. Ludtke, Yifei Kong, Salih J. Wakil, Jianpeng Ma, and Wah Chiu. Experimental verification of conformational variation of human fatty acid synthase as predicted by normal mode analysis. *Structure*, 12:185–191, 2004.
- [15] S.K. Brumfield, A.C. Ortmann, V. Ruigrok, P. Suci, T. Douglas, and M.J. Young. Particle assembly and ultrastructural features associated with replication of the lytic archaeal virus *sulfolobus turreted icosahedral virus*. *J. Virol.*, 83:5964–5970, 2009.
- [16] James R. Bunch, Christopher P. Nielsen, and Danny Sorensen. Rank-one modification of the symmetric eigenproblem. *Numer. Math.*, 31:31–48, 1978.
- [17] S.J. Butcher, D.H. Bamford, and S.D Fuller. DNA packaging orders the membrane of bacteriophage PRD1. *EMBO J.*, 14:6078–6086, 1995.
- [18] Giovanni Cardone, Dennis C. Winkler, Benes L. Trus, Naiqian Cheng, John E. Heuser, William W. Newcomb, Brown Jay C., and Alasdair C. Steven. Visualization of the herpes simplex virus portal *in situ* by cryo-electron tomography. *Journal of Virology*, 361:426–434, 2007.
- [19] S. Casjens and J. King. Virus assembly. *Annu. Rev. Biochem.*, 44:555–611, 1975.
- [20] Juan T. Chang, Michael Schmid, Frazer J. Rixon, and Wah Chiu. Electron cryotomography reveals the portal in the herpesvirus capsid. *Journal of Virology*, 81:2065–2068, 2007.
- [21] Earl A. Coddington and Norman Levinson. *Theory of Ordinary Differential Equations*. McGraw-Hill, New York, 1955.

- [22] Thomas S. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.
- [23] Binbin Deng, Christine M. O'Connor, Dean H. Kedes, and Z. Hong Zhou. Direct visualization of the putative portal in the kaposi's sarcoma-associated herpesvirus capsid by cryoelectron tomography. *Journal of Virology*, 81:3640–3644, 2007.
- [24] Peter C. Doerschuk and John E. Johnson. *Ab initio* reconstruction and experimental design for cryo electron microscopy. *IEEE Trans. Info. Theory*, 46(5):1714–1729, August 2000. Digital Object Identifier 10.1109/18.857786.
- [25] T. Dokland and H. Murialdo. Structural transitions during maturation of bacteriophage lambda capsids. *J. Mol. Biol.*, 233:682–694, 1993.
- [26] J. Dubochet, M. Adrian, J.J. Chang, J.C. Homo, J. Lepault, A.W. McDowell, and P. Schultz. Cryo-electron microscopy of vitrified specimens. *Q Rev. Biophys.*, 21:129–228, 1988.
- [27] H. Engelhardt and J. Peters. Structural research on surface layers: a focus on stability, surface layer homology domains, and surface layer-cell wall interactions. *J. Struct. Biol.*, 124:276–302, 1998.
- [28] A. Erdelyi, editor. *Higher Transcendental Functions*. McGraw-Hill, 1953.
- [29] Burak Erman. The Gaussian network model: Precise prediction of residue fluctuations and application to binding problems. *Biophys. J.*, 91:3589–3599, 2006.
- [30] Friedrich Forster, Ohad Medalia, Nathan Zauberman, Wolfgang Baumeister, and Deborah Fass. Retrovirus envelope protein complex structure *in situ* studied by cryo-electron tomography. *Proc. Nat. Acad. Sci. U.S.A.*, 102(13):4729–4734, 2004.
- [31] Friedrich Forster, Sabine Pruggnaller, Anja Seybert, and Achilleas S. Frangakis. Classification of cryo-electron sub-tomograms using constrained correlation. *Journal of Structural Biology*, 161:276–286, 2008.
- [32] A.S. Frangakis, J. Bohm, F. Forster, S. Nickell, D. Nicastro, D. Typke, R. Hegerl, and W. Baumeister. Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. *Proc. Nat. Acad. Sci. U.S.A.*, 99:14153–14158, 2002.

- [33] Joachim Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Academic Press, San Diego, 1996.
- [34] Joachim Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*. Oxford University Press, Oxford, 2 edition, 2006.
- [35] Chi-yu Fu, Kang Wang, Jason Lanman, Reza Khayat, Mark J. Young, Grant J. Jensen, Peter C. Doerschuk, and John E. Johnson. In vivo assembly of an archaeal virus studied with whole cell electron cryotomography. *Structure*, 18(12):1579–1586, 2010.
- [36] D.N. Fuller, D.M. Raymer, V.I. Kottadiel, V.B. Rao, and D.E. Smith. Single phage t4 dna packaging motors exhibit large force generation, high velocity, and dynamic variability. *Proc. Nat. Acad. Sci. U.S.A.*, 104:16868–16873, 2007.
- [37] G. Gaietta, T.J. Deerinck, S.R. Adams, J. Bouwer, O. Tour, D.W. Laird, G.E. Sosinsky, R.Y. Tsien, and M.H. Ellisman. Multicolor and electron microscopic imaging of connexin trafficking. *Science*, 296:503–507, 2002.
- [38] L. Gan, J. A. Speir, J. F. Conway, G. Lander, N. Cheng, B. A. Firek, R. W. Hendrix, R. L. Duda, L. Liljas, and J. E. Johnson. Capsid conformational sampling in HK97 maturation visualized by x-ray crystallography and cryo-EM. *Structure*, 14(11):1655–1665, November 2006.
- [39] W.M. Gelbart and C.M. Knobler. Virology. Pressurized viruses. *Science*, 323:1682–1683, 2009.
- [40] G. H. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM J. Numer. Anal.*, 10(2):413–432, April 1973.
- [41] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 3rd edition, 1996.
- [42] B. Gowen, J.K. Bamford, D.H. Bamford, and S.D Fuller. The tailless icosahedral membrane virus prd1 localizes the proteins involved in genome packaging and injection at a unique vertex. *J. Virol.*, 77:7863–7871, 2003.
- [43] Nikolaus Grigorieff and Stephen C. Harrison. Near-atomic resolution

- reconstructions of icosahedral viruses from electron cryo-microscopy. *J. Struct. Biol.*, 21:265–273, 2011.
- [44] R. Grimm, H. Singh, R. Rachel, D. Typke, W. Zillig, and W. Baumeister. Electron tomography of ice-embedded prokaryotic cells. *Biophys. J.*, 74:1031–1042, 1998.
  - [45] K. Grunewald and M. Cyrklaff. Structure of complex viruses and virus-infected cells by electron cryo tomography. *Curr. Opin. Microbiol.*, 9:437–442, 2006.
  - [46] Kay Grunewald, Prashant Desai, Dennis C. Winkler, J. Bernard Heymann, David M. Belnap, Wolfgang Baumeister, and Alasdair C. Steven. Three-dimensional structure of herpes simplex virus from cryo-electron tomography. *Science*, 302:1396–1398, 2003.
  - [47] D. M. Healy, D. N. Rockmore, P. J. Kostelec, and S. Moore. Ffts for the 2-sphere-improvements and variations. *The Journal of Fourier Analysis and Applications*, 9:341–385, 2003.
  - [48] G.P. Henderson, L. Gan, and G.J. Jensen. 3-d ultrastructure of o. tauri: electron cryotomography of an entire eukaryotic cell. *PLoS One*, 2:e749, 2007.
  - [49] Kerson Huang. *Statistical Mechanics*. John Wiley and Sons, Inc., New York, 2 edition, 1987.
  - [50] Juha T. Huiskonen, Jussi Hepojoki, Pasi Laurinmaki, Antti Vaheri, Hilikka Lankinen, Sarah J. Butcher, and Kay Grunewald. Electron cryotomography of tula hantavirus suggests a unique assembly paradigm for enveloped viruses. *Journal of Virology*, 84(10):4889–4897, 2010.
  - [51] C. V. Iancu, W. F. Tivol, J. B. Schooler, D. P. Dias, G. P. Henderson, G. E. Murphy, E. R. Wright, Z. Li, Z. Yu, and A. Briegel. Electron cryotomography sample preparation using the vitrobot. *Nat. Protoc.*, 1:2813–2819, 2006.
  - [52] Kenji Iwasaki and Toshihiro Omura. Electron tomography of the supramolecular structures of virus-infected cells. *J. Struct. Biol.*, 20:632–639, 2010.

- [53] John David Jackson. *Classical Electrodynamics*. John Wiley, New York, 2nd edition, 1975.
- [54] Wen Jiang, Juan Chang, Joanita Jakana, Peter Weigele, Jonathan King, and Wah Chiu. Structure of epsilon15 bacteriophage reveals genome organization and DNA packaging/injection apparatus. *Nature*, 439(7076):612–616, February 2 2006.
- [55] J. E. Johnson. Virus particle dynamics. *Adv. Protein Chem.*, 64:197–218, 2003.
- [56] J.E. Johnson and W. Chiu. DNA packaging and delivery machines in tailed bacteriophages. *Curr. Opin. Struct. Biol*, 17:237–243, 2007.
- [57] Linda Kaufman. A variable projection method for solving separable non-linear least squares problems. *BIT*, 15:49–57, 1975.
- [58] Reza Khayat, Liang Tang, Eric T. Larson, C. Martin Lawrence, Mark Young, and John E. Johnson. Structure of an archaeal virus capsid protein reveals a common ancestry to eukaryotic and bacterial viruses. *Proc. Nat. Acad. Sci. U.S.A.*, 102(52):18944–18949, 2005.
- [59] K. Kim, Moon, Gregory S. Cirikjian, and Robert L. Jernigan. Elastic models of conformational transitions in macromolecules. *Journal of Molecular Graphics and Modelling*, 21:151–160, 2002.
- [60] K. Kim, Moon, Robert L. Jernigan, and Gregory S. Cirikjian. Efficient generation of feasible pathways for protein conformational transitions. *Biophysical Journal*, 83:1620–1630, 2002.
- [61] K. Kim, Moon, Robert L. Jernigan, and Gregory S. Cirikjian. Elastic models of conformational transitions in macromolecules. *J. Struct. Biol.*, 143:107–117, 2003.
- [62] A. Komeili, Z. Li, D.K. Newman, and G.J. Jensen. Magnetosomes are cell membrane invaginations organized by the actin-like protein MamK. *Science*, 311:242–245, 2006.
- [63] Peter J. Kostelec and Daniel N. Rockmore. Ffts on the rotation group. *The Journal of Fourier Analysis and Applications*, 14:145–179, 2008.
- [64] A.J. Koster, R. Grimm, D. Typke, R. Hegerl, A. Stoschek, J. Walz, and

- W. Baumeister. Perspectives of molecular and cellular electron tomography. *J. Struct. Biol.*, 120:276–308, 1997.
- [65] James R. Kremer, David N. Mastronarde, and J. Richard McIntosh. Computer visualization of three-dimensional image data using IMOD. *Journal of Structural Biology*, 116:71–76, 1996.
- [66] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951.
- [67] J. Kurner, A.S. Frangakis, and W. Baumeister. Cryo-electron tomography reveals the cytoskeletal structure of *spiroplasma melliferum*. *Science*, 307:436–438, 2005.
- [68] Peter Lancaster and Miron Tismenetsky. *The Theory of Matrices*. Academic Press, 2 edition, 1985.
- [69] Gabriel C. Lander, Liang Tang, Sherwood R. Casjens, Eddie B Gilcrease, Peter Prevelige, Anton Poliakov, Clinton S. Potter, Bridget Carragher, and John E. Johnson. The structure of an infectious P22 virion shows the signal for headful DNA packaging. *Science*, 312(5781):1791–1795, 23 June 2006.
- [70] G.C. Lander, A. Evilevitch, M. Jeembaeva, C.S. Potter, B. Carragher, and J.E Johnson. Bacteriophage lambda stabilization by auxiliary protein gpD: timing, location, and mechanism of attachment determined by cryo-em. *Structure*, 16:1399–1406, 2008.
- [71] J. Lanman, J. Crum, T. J. Deerinck, G. M. Gaietta, A. Schneemann, G. E. Sosinsky, M. H. Ellisman, and J. E. Johnson. Visualizing flock house virus infection in drosophila cells with correlated fluorescence and electron microscopy. *J. Struct. Biol.*, 161:439–446, 2008.
- [72] Junghoon Lee, Peter C. Doerschuk, and John E. Johnson. Exact reduced-complexity maximum likelihood reconstruction of multiple 3-D objects from unlabeled unoriented 2-D projections and electron microscopy of viruses. *IEEE Trans. Image Proc.*, 16(11):2865–2878, November 2007. PMID: 18092587.
- [73] Junghoon Lee, Zhye Yin, Peter C. Doerschuk, Jinghua Tang, and John E. Johnson. Automatic *ab initio* simultaneous classification and 3-D reconstruction of multiple types of viruses from cryo electron microscope images showing a mixture of all types. Submitted to *J. Struct. Biol.*

- [74] Andrew P. Leis, Martin Beck, Manuela Gruska, Christoph Best, Reiner Hegerl, Wolfgang Baumeister, and John W. Leis. Cryo-electron tomography of biological specimens. *IEEE Sig. Proc. Mag.*, 23(3):95–103, 2006.
- [75] V. Lucic, F. Forster, and W. Baumeister. Structural studies by electron tomography: from cells to molecules. *Annu. Rev. Biochem.*, 74:833–865, 2005.
- [76] Vladan Lucic, Friedrich Forster, and Wolfgang Baumeister. Structural studies by electron tomography: From cells to molecules. *Annual Review of Biochemistry*, 74:833–865, 2005.
- [77] W. S. Maaty, A. C. Ortmann, M. Dlakic, K. Schulstad, J. K. Hilmer, L. Liepold, B. Weidenheft, R. Khayat, T. Douglas, M. J. Young, and B Bothner. Characterization of the archaeal thermophile *Sulfolobus* turreted icosahedral virus validates an evolutionary link among double-stranded DNA viruses from all domains of life. *J. Virol.*, 80:7625–7635, 2006.
- [78] M. Marko, C. Hsieh, R. Schalek, J. Frank, and C Mannella. Focused-ion-beam thinning of frozen-hydrated biological specimens for cryo-electron microscopy. *Nat. Methods.*, 4:215–217, 2007.
- [79] C. S. Martin, R. M. Burnett, F. de Haas, R. Heinkel, T. Rutten, S. D. Fuller, S. J. Butcher, and D. H Bamford. Combined em/x-ray imaging yields a quasi-atomic model of the adenovirus-related bacteriophage PRD1 and shows key capsid and membrane interactions. *Structure*, 9:917–930, 2001.
- [80] David N Mastronarde. Dual-axis tomography: an approach with alignment methods that preserve resolution. *Journal of Structural Biology*, 120:343–352, 1997.
- [81] O. Medalia, M. Beck, M. Ecke, I. Weber, R. Neujahr, W. Baumeister, and G Gerisch. Organization of actin networks in intact filopodia. *Curr. Biol.*, 17:79–84, 2007.
- [82] O. Medalia, I. Weber, A. S. Frangakis, D. Nicastro, G. Gerisch, and W Baumeister. Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science*, 298:1209–1213, 2002.
- [83] Dengming Ming, Yifei Kong, Maxime A. Lambert, Zhong Huang, and Jianpeng Ma. How to describe protein motion without amino acid sequence and atomic coordinates. *Proc. Nat. Acad. Sci. U.S.A.*, 99(13):8620–8625, 25 June 2002.

- [84] S. D. Moore and Jr Prevelige, P. E. DNA packaging: a new class of molecular motors. *Curr. Biol.*, 12:R96–98, 2002.
- [85] J. O. Ortiz, F. Forster, J. Kurner, A. A. Linaroudis, and W Baumeister. Mapping 70S ribosomes in intact cells by cryoelectron tomography and pattern recognition. *J. Struct. Biol.*, 156:334–341, 2006.
- [86] A. C. Ortmann, B. Wiedenheft, T. Douglas, and M. Young. Hot crenarchaeal viruses reveal deep evolutionary connections. *Nat. Rev. Microbiol.*, 4:520–528, 2006.
- [87] P. Ostapchuk and P Hearing. Control of adenovirus packaging. *J. Cell. Biochem.*, 96:25–35, 2005.
- [88] Pawel Penczek, Michael Radermacher, and Joachim Frank. Three-dimensional reconstruction of single particles embedded in ice. *Ultra-microscopy*, 40:33–53, 1992.
- [89] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.*, 25(13):1605–1612, 2004.
- [90] Jr Prevelige, P.E. Send for reinforcements! Conserved binding of capsid decoration proteins. *Structure*, 16:1292–1293, 2008.
- [91] Cory J. Prust, Peter C. Doerschuk, Gabriel C. Lander, and John E. Johnson. *Ab initio* maximum likelihood reconstruction from cryo electron microscopy images of an infectious virion of the tailed bacteriophage P22 and maximum likelihood versions of Fourier Shell Correlation appropriate for measuring resolution of spherical or cylindrical objects. *J. Struct. Biol.*, 167:185–199, 2009. PubMedCentral: PMC1803348, PMID: 19457456, NIHMSID: 119275, Digital Object Identifier 10.1016/j.jsb.2009.04.013.
- [92] A.J. Rader, Chakra Chennubhotla, Lee-Wei Yang, and Ivet Bahar. The gaussian network model: Theory and applications. In Qiang Cui and Ivan Bahar, editors, *Normal Mode Analysis. Theory and Applications to Biological and Chemical Systems*, chapter 3, pages 41–64. Chapman and Hall, 2006.
- [93] G. Rice, L. Tang, K. Stedman, F. Roberto, J. Spuhler, E. Gillitzer, J. E. Johnson, T. Douglas, and M. Young. The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. *Proc. Nat. Acad. Sci. U.S.A.*, 101:7716–7720, 2004.



- [94] W. H. Roos, I. L. Ivanovska, A. Evilevitch, and G. J. Wuite. Viral capsids: mechanical characteristics, genome packaging and delivery mechanisms. *Cell. Mol. Life Sci.*, 64:1484–1497, 2007.
- [95] Morris E. Rose. *Elementary Theory of Angular Momentum*. John Wiley and Sons, New York, 1957.
- [96] S. H. Scheres, R. Nunez-Ramirez, Y. Gomez-Llorente, C. San Martin, P. P. Eggermont, and J. M. Carazo. Modeling experimental image formation for likelihood-based classification of electron microscopy data. *Structure*, 15:1167–1177, 2007.
- [97] S. H. Scheres, M. Valle, R. Nunez, Sorzano C. O., R. Marabini, G. T. Herman, and J. M Carazo. Maximum-likelihood multi-reference refinement for electron microscopy images. *J. Mol. Biol.*, 348:139–149, 2005.
- [98] Sjors H. W. Scheres, Haixiao Gao, Mikel Valle, Gabor T. Herman, Paul P. B. Eggermont, Joachim Frank, and Jose-Maria Carazo. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature Methods*, 4(1):27–29, January 2007.
- [99] Sjors H. W. Scheres, Roberto Melero, Mikel Valle, and Jose-Maria Carazo. Averaging of electron subtomograms and random conical tilt reconstructions through likelihood optimization. *Structure*, 17:1563–1572, 9 December 2009.
- [100] Michael F. Schmid and Christopher R. Booth. Methods for aligning and for averaging 3D volumes with missing data. *Journal of Structural Biology*, 161:243–248, 2008.
- [101] D.E. Smith, S.J. Tans, S.B. Smith, S. Grimes, D.L. Anderson, and C. Bustamante. The bacteriophage straight phi29 portal motor can package dna against a large internal force. *Nature*, 413:748–752, 2001.
- [102] J. C. Snyder, K. Stedman, G. Rice, B. Wiedenheft, J. Spuhler, and M. J. Young. Viruses of hyperthermophilic archaea. *Res. Microbiol.*, 154:474–482, 2003.
- [103] Adam E. Sougrat, Alberto Bartesaghi, Jeffrey D. Lifson, Adam E. Bennett, Julian W. Bess, Daniel J. Zabransky, and Sriam Subramaniam. Electron tomography of the contact between t cells and siv/hiv-1:implications for viral entry. *PLoS Pathogens*, 3:e63, 2007.

- [104] Sriram Subramaniam, Alberto Bartesaghi, Jun Liu, Adam E. Bennet, and Rachid. Sougrat. Electron tomography of viruses. *J. Struct. Biol.*, 17:596–602, 2007.
- [105] C. Suloway, J. Pulokas, D. Fellmann, A. Cheng, F. Guerra, J. Quispe, S. Stagg, C. S. Potter, and B Carragher. Automated molecular microscopy: the new leginon system. *J. Struct. Biol.*, 151:41–60, 2005.
- [106] C. Suloway, J. Shi, A. Cheng, J. Pulokas, B. Carragher, C. S. Potter, S. Q. Zheng, D. A. Agard, and G. J Jensen. Fully automated, sequential tilt-series acquisition with leginon. *J. Struct. Biol.*, 167:11–18, 2009.
- [107] F. Tama, W. Wriggers, and C. L. Brooks III. Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. *J. Mol. Biol.*, 321:297, 2002.
- [108] Florence Tama, Osamu Miyashita, and Charles L. Brooks, III. Normal model based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J. Struct. Biol.*, 147:315–326, 2004.
- [109] C. S. Ting, C. Hsieh, S. Sundararaman, C. Mannella, and M Marko. Cryo-electron tomography reveals the comparative three-dimensional architecture of Prochlorococcus, a globally important marine cyanobacterium. *J. Bacteriol.*, 189:4485–4493, 2007.
- [110] M. M. Tirion. Large amplitude elastic motions in proteins from a single-parameter atomic analysis. *Phys. Rev. Lett.*, 77:1905–1908, 1996.
- [111] Marin van Heel, Brent Gowen, Rishi Matadeen, Elena V. Orlova, Robert Finn, Tillmann Pape, Dana Cohen, Holger Stark, Ralf Schmidt, Michael Schatz, and Ardan Patwardhan. Single-particle electron cryo-microscopy: towards atomic resolution. *Quarterly Reviews of Biophysics*, 33(4):307–369, 2000.
- [112] J. Walz, D. Typke, M. Nitsch, A. J. Koster, R. Hegerl, and W Baumeister. Electron tomography of single ice-embedded macromolecules: Three-dimensional alignment and classification. *J. Struct. Biol.*, 120:387–395, 1997.
- [113] Kang Wang, Chi-yu Fu, Peter C. Doerschuk, and John E. Johnson. *In vivo* virus structures: Simultaneous classification, resolution enhancement,

- and noise reduction in whole-cell electron tomography. *J. Struct. Biol.*, 174(3):425–433, 2011.
- [114] Tianyun Wei, Tamaki Uehara-Ichiki, Naoyuki Miyazaki, Hiroyuki Hibino, Kenji Iwasaki, and Toshihiro Omura. Association of rice gall dwarf virus with microtubules is necessary for viral release from cultured insect vector cells. *jvlong*, 83:10830–10835, 2009.
- [115] William R. Wikoff, James F. Conway, Jinghua Tang, Kelly K. Lee, Lu Gan, Naiqian Cheng, Robert L. Duda, Roger W. Hendrix, Alasdair C. Steven, and John E. Johnson. Time-resolved molecular dynamics of bacteriophage HK97 capsid maturation interpreted by electron cryo-microscopy and x-ray crystallography. *J. Struct. Biol.*, 153:300–306, 2006.
- [116] Hanspeter Winkler, Ping Zhu, Jun Liu, Feng Ye, Kenneth H. Roux, and Kenneth A. Taylor. Tomographic subvolume alignment and subvolume classification applied to myosin V and SIV envelope spikes. *Journal of Structural Biology*, 165:64–77, 2009.
- [117] Lee-Wei Yang and Ivet Bahar. Coupling between catalytic site and collective dynamics: A requirement for mechanochemical activity of enzymes. *Structure*, 13:893–904, 2005.
- [118] Zhye Yin, Peter C. Doerschuk, and Saul B. Gelfand. Model calculations for joint pattern recognition and signal reconstruction in cryo electron microscopy. *Communications in Information and Systems*, 4(1):73–88, 2004. Special Issue in honor of the 70th birthday of Professor Sanjoy K. Mitter, online at [http://www.ims.cuhk.edu.hk/~cis/2004.1/yin\\_etal.pdf](http://www.ims.cuhk.edu.hk/~cis/2004.1/yin_etal.pdf).
- [119] Zhye Yin, Yili Zheng, Peter C. Doerschuk, Padmaja Natarajan, and John E. Johnson. A statistical approach to computer processing of cryo electron microscope images: Virion classification and 3-D reconstruction. *J. Struct. Biol.*, 144(1/2):24–50, 2003. PMID: 14643207.
- [120] Wei Zhang, Marek Kimmel, Christian M. T. Spahn, and Pawel A. Penczek. Heterogeneity of large macromolecular complexes revealed by 3D cryo-EM variance analysis. *Structure*, 16:1770–1776, 2008.
- [121] Yibin Zheng and Peter C. Doerschuk. Explicit orthonormal fixed bases for spaces of functions that are totally symmetric under the rotational symmetries of a Platonic solid. *Acta Cryst.*, A52:221–235, 1996.

- [122] Yibin Zheng and Peter C. Doerschuk. 3D image reconstruction from averaged Fourier transform magnitude by parameter estimation. *IEEE Trans. Image Proc.*, 7(11):1561–1570, November 1998.
- [123] Yibin Zheng and Peter C. Doerschuk. Explicit computation of orthonormal symmetrized harmonics with application to the identity representation of the icosahedral group. *SIAM Journal on Mathematical Analysis*, 32(3):538–554, 2000.
- [124] Yili Zheng. *Novel statistical models and a high-performance computing toolkit for the solution of cryo electron microscopy inverse problems in viral structural biology*. PhD thesis, School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA, August 2008.
- [125] Yili Zheng and Peter C. Doerschuk. 3-D signal reconstruction as a generalization of Gaussian mixture parameter estimation. *IEEE Trans. Image Proc.* In review.
- [126] Yili Zheng and Peter C. Doerschuk. A parallel software toolkit for statistical 3-D virus reconstructions from cryo electron microscopy images using computer clusters with multi-core shared-memory nodes. In *22nd IEEE International Parallel and Distributed Processing Symposium (IPDPS 2008)*, pages 1–11, Miami, Florida, April 14–18 2008. IEEE. Digital Object Identifier 10.1109/IPDPS.2008.4536242.
- [127] Yili Zheng and Peter C. Doerschuk. Stochastic 3-d signal reconstruction from noisy projection data for heterogeneous instances of objects in electron microscopy imagery. In *2011 IEEE International Symposium on Biomedical Imaging (ISBI'11)*, pages 918–921, March 29–July 2 2011. DOI 10.1109/ISBI.2011.5872630.
- [128] Ping Zhu, Hanspeter Winkler, Elena Chertova, Kenneth A. Taylor, and Kenneth H. Roux. Cryoelectron tomography of hiv-1 envelope spikes: Further evidence for tripod-like legs. *PLoS Pathogens*, 4(11):e1000203, 2008.
- [129] G. Ziedaite, H.M. Kivela, J.K. Bamford, and D.H. Bamford. Purified membrane-containing procapsids of bacteriophage prd1 package the viral genome. *J. Mol. Biol.*, 386:637–647, 2009.